DOCUMENT RESUME

ED 213 727 TM 820 053

AUTHOR Smith, Laura Spooner
TITLE Investigation of Writing Assessment Strategies. Studies in Measurement & Methodology. Work Unit 1: Design and Use of Tests.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Nov 78
GRANT OB-NIE-G-78-0213
NOTE 100p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Evaluation Methods; High Schools; *High School Students; *Rating Scales; Scores; Writing (Composition); *Writing Evaluation
IDENTIFIERS *Analytic Rating Procedures; *Impressionistic Rating Procedures

ABSTRACT

Measurement experts in the field of writing and writing-related competencies have to deal with fluctuating test reliability, implementation of efficient scoring procedures, and examining the validity of objective measures of writing. This study was undertaken to address measurement concerns brought to light by the competency-based assessment movement. The primary aim of the study was to examine relationships among three writing assessment strategies which are potentially useful within the context of competency-based testing. Two of the strategies can be classified as direct measures of performance, involving collecting and rating actual samples of student writing. The third strategy, an indirect measure of performance, involved administration of an objective test of writing skills presumably linked to the two direct measure strategies. Implications of the findings for both instructional and measurement practices are presented. (Author/CE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED213727

DELIVERABLE: NOVEMBER 1978
Contract #OB-NIEG-78-0213

STUDIES IN MEASUREMENT & METHODOLOGY

Work Unit 1:
Design and Use of Tests

Eva L. Baker
Project Director

INVESTIGATION OF WRITING ASSESSMENT STRATEGIES

Laura Spooner Smith

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G.S. Gray
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 053

Center for the Study of Evaluation
Graduate School of Education
University of California - Los Angeles

The work upon which this publication is based was performed pursuant to a grant with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated to not necessarily represent official NIE position or policy.

TABLE OF CONTENTS

I. Introduction 1
The Problem 1
The Response Mode Issue 2
The Response Criteria Issue 3
Purposes and Research Questions 5

II. Literature Review 7
Assessment of Writing and Writing-Related Competencies 7
Research on the Reliability of Writing Samples 7
The Writer Variable 8
The Assignment Variable 9
The Rater Variable 10
The Colleague Variable 11
Guided Scoring Procedures 13
The Contribution of Indirect Measures of Writing 15
An Analytic Strategy: The College English Placement Test 16
Empirical Approaches: The ETS English Composition Test 18
Competency-based Assessment and the Selection of Writing Tasks 21

III: Method 23
Method -- Overview 23
The Sample 23
Treatment Groups 25
Study Procedures -- Overview 25
Measures 27
The Writing Task 27
The Writing Stimulus (Topics and Directions) 27
The Response Criteria 31
The Analytic Scale 31
The Impressionistic Scale 38
The Objective Item Measure 39

TABLE OF CONTENTS (Continued)

Rating Procedures 48
Impressionistic Scale Rater Training 49
Impressionistic Scale Rating Session 53
Basic Measurement Properties of Impressionistic Scale 54
Analytic Scale Rater Training 54
Analytic Scale Rating Session 57
Basic Measurement Properties of Analytic Scale 59
Objective Item Test Procedures 61
Basic Measurement Properties of Objective Item Test 61

IV: Findings. 68
Findings -- Overview 68
Relationships Among Analytic and Impressionistic Scale Scores 68
Relationships Among Rating Scale Scores and the Objective Measure . . 71
Summary of Findings 76

V: Discussion 79
Overview 79
Discussion of Findings 80
Relationships Among the Direct Measures of Writing 80
Relationships Between the Indirect and Direct Measures of Writing 83
Implications for Educational Practice 86

REFERENCES 89

FIGURES

1. Measurement Strategies 6
2. Treatment Groups Within Each Classroom 25
3. Abbreviated Version of the Analytic Rating Scale 33
4. Relationship Between Analytic Scale Elements and Objective Test Items 42
5. Forms of the Objective Measure 41
6. Descriptors for Six Point Impressionistic Rating Scale 50
7. Design of Impressionistic and Analytic Rating Procedures 52

TABLES

1. Teacher Estimate of Participating Classes' Writing Ability 24
2. Means and Standard Deviations on Impressionistic Scale Scores 55
3. Rater Reliabilities of Analytic Scale Pilot Test 58
4. Rater Reliabilities of Analytic Scale Final Study 59
5. Means and Standard Deviations on Analytic Scales 60
6. Means and Standard Deviations on Total Objective Test and Subsections for the Total Population 62
7. Means and Standard Deviations on Total Objective Test and Subsections by Class 64
8. Objective Measure Coefficients of Reliability 63
9. Objective Measure Subsection VI Stem Correlations 65
10. Intercorrelations Among Impressionistic and Analytic Rating Scale Scores 69
11. Regression of Impressionistic Scale Scores on Six Analytic Subscales . 72
12. Regression of Impressionistic Scale Scores on Three Analytic Subscales 72

TABLES (continued)

13. Correlations Among Analytic and Impressionistic Total Scores and Objective Measure Scores 74

14. Regression of Analytic Scale Total on Six Objective Measure Subtests 75

15. Regression of Impressionistic Scale Total on Six Objective Measure Subtests 75

16. Regression of Analytic Subscales on Analogous Objective Measure Subtests 77

APPENDICES

A. Notice to Participating Teachers

B. Data Collector Script

C. Guidelines for Developing Writing Sample Topics

D. Final Study Topics

E. Pilot Test Topics

F. The Analytic Rating Scale

G. Condensed Version of Analytic Scale

H. Objective Measure Item Generation Guidelines

I. Objective Item Measure

J. Impressionistic Rating Scale Rater Training Materials

K. Variance Components for Impressionistic Rating Scale Pilot Test

L. Variance Components for Impressionistic Rating Final Study

M. Variance Components for Analytic Rating Pilot Test

N. Variance Components for Analytic Scale Rating Final Study

INVESTIGATION OF WRITING ASSESSMENT STRATEGIES

INTRODUCTION

The Problem

As demand increases for ways to assess students' basic academic skills, additional requirements for measures of writing proficiency also have surfaced. The need today is for measures of writing that not only are technically sound, but which also promise to yield unambiguous information about students' writing skill deficits and strengths. This situation has prompted measurement and curriculum specialists to seriously re-examine state-of-the-art approaches to writing assessment and also has encouraged fresh inquiries into time honored questions about the nature of written discourse and its evaluation. For example: What testing approaches will provide clear information about attainment of specific writing skills? What skills, in fact, constitute the "domains" of writing? What measurement strategies are amenable to economic scoring while still providing maximally useful information to public officials, parents and educational decision makers?

Such questions reflect a desire on the part of educational practitioners to develop and use measures that provide both useful and useable information about their students' writing skills. Moreover, the questions have directed renewed attention to a fundamental concern in writing assessment: how to identify and define measurement tasks that will serve as meaningful and efficient indicators of writing competence.

Determining meaningful test tasks (or items) to measure a domain of learning is an intellectually difficult aspect of competency-based assessment. As Glaser and Resnick (1972) have pointed out, determining critical attributes

of what is learned is not a well developed science, especially for complex behaviors such as the production of written discourse. Millman's (1974) assertion that a performance domain should be defined"...by those facets and elements that make a difference in how the learner responds" lead naturally to the question, what facets make a difference?

The Response Mode Issues. In the measurement of writing skills, one "facet" that clearly makes a difference is the form of response that ought to be elicited from examinees. Indeed, an historically controversional issue in writing assessment centers on the relative merits of two contending response modes: production of a piece of writing, i.e., "writing samples," and selecting a response from among given alternatives, i.e., "objective" tests of writing.

That there is any question regarding the most appropriate approach for assessing composition skills at first may appear puzzling. Clearly, writing samples represent a direct measure of writing performance and therefore possess prima facie validity. This, in fact, is the principal argument supplied by proponents of writing samples as a means of composition assessment (Coffman, 1971). Those who favor more indirect methods, i.e., objective tests, however, have pointed to the usual unreliability of writing samples, notably the difficulty of ensuring reliable scoring procedures and the tendency of a writer's performance to fluctuate in quality over time (Yeasmeen and Barker, 1973). Problems of unreliability in writing samples have been the subject of recent research which has resulted in promising procedures aimed at enhancing measurement reliability. Nonetheless, a clear advantage of objective methods of assessment is the efficiency with which the measures can be administered

and scored, an economy of especial importance when sizeable aggregates of students are to be tested.

Arguments of efficiency, though, have failed to impress many professionals in the discipline of English (e.g., Cooper and Odell, 1977), who have voiced concern over the inherent lack of content validity of indirect approaches to writing assessment. In response to such objections, proponents of indirect measures have cited findings from studies (e.g., Godshalk, 1966) which reveal statistically significant correlations between performance on objective items intended to assess aspects of writing competence and performance on actual writing samples.

Regardless of such positive characteristics, objective measures of writing remain indirect predictors of students' ability to produce actual samples of writing. To date, there are no conclusive data to indicate that the "writing" skills presumably tapped by objective measures bear a clear and direct relationship to production of written discourse. At best, findings from correlational studies indicate that various combinations of skills measured by objective tests are related in an unspecified fashion to proficiency in writing. Which components of writing ability, if any, such objective measures do predict remains a matter of conjecture. More precise evidence regarding the class of skills measured by objective items is needed if such indirect strategies are to be of value within the context of competency based assessment of writing.

The Response Criteria Issue. There is yet another issue at stake in the selection of tasks to measure writing competence, one that applies especially to direct methods of assessment. When the stimulus attributes of a task

require students to produce a piece of writing, the means by which the writing is rated becomes critically important. Indeed, attributes of the testing stimulus and attributes of the rating criteria are integrally and iteratively linked; together they constitute an operational definition of the entire measurement task. Concerns unrelated to the shaping of student performances, but at the heart of judgments regarding adequacy of the performance, thus arise. For example, do different strategies for rating the same sample of writing systematically provide different estimates of competence? In addition to questions about the consistency of judgments across rating schemes, the current emphasis on competency based assessment directs attention to the type of information yielded by rating schemes. For instance, do some rating approaches paint a more complete picture of skill strengths and deficits than other approaches? Can such information be obtained through a reasonable outlay of resources, both time and money?

Although a variety of guided procedures for sorting or ranking written pieces currently are in use, most can be classified as either "analytic" or "holistic." Analytic procedures require raters to assign points to each of several specified aspects of a composition and yield an estimate of overall quality of the writing product, as well as subscores on individual rhetorical and structural features of the writing sample. In contrast, holistic rating procedures require raters to assign a single value to a piece of writing. Reliability of ratings typically is achieved through one of two different procedures. The "norming session" procedure requires raters to compare samples of writing to a carefully selected set of compositions ranging in quality and to assign a score according to the correspondence between the

sample and the "norming" compositions. Another procedure provides raters with a written rubric outlining levels of quality. The score assigned by the rater is intended to reflect the level of quality exhibited by the writing sample. Impressionist scoring procedures also require assignment of a single value to a written product. Unlike other holistic approaches, impressionistic procedures involve only a minimal rubric to guide the judgments of raters.

The apparent advantage of holistic rating procedures is that they tend to be an efficient and reliable method for obtaining judgments on the overall quality of a written product. However, analytic approaches make clear and public the criteria of evaluation, a characteristic which is especially important when diagnostic data and/or instructional remediation are required. Whether analytic and holistic approaches provide substantially different estimates of competence, however, is an issue requiring further study.

Purposes and Research Questions

The overall purpose of this study was to explore the relationships among alternate strategies to measure high school students' performance within a defined domain of writing skill. Three measurement strategies were employed and are depicted in Figure 1.

Figure 1

MEASUREMENT STRATEGIES

Form of Response:	
I	II
Production	Selection
A. Writing samples judged by Analytic scoring criteria	A. Objective items requiring students to discriminate among given passages of written discourse.
B. Writing samples judged by Impressionistic scoring criteria	

Research questions addressed by the study were concerned with the relationships among students' performances across the varying writing tasks. Specifically, questions addressed were as follows:

1. What is the relationship between Analytic and Impressionistic rating scale scores?
2. What is the relationship between Analytic rating scale scores and scores on the Objective items?
3. What is the relationship between Impressionistic rating scale scores and scores on the Objective items?

LITERATURE REVIEW

Assessment of Writing and Writing-Related Competencies

The measurement of writing and writing-related competencies historically has been a problematic and controversial venture. One issue that has captured the attention of test developers and professionals in English concerns the form of response that should be elicited from examinees. The issue has focussed on two contending forms of response: production of a piece of writing, i.e., writing samples, and selecting a response from given alternatives, i.e., objective items. Much of the discussion has centered on the relative validities and measurement reliability of the two forms of response.

Research on the reliability of writing samples. That writing samples as tests of composition skill have face validity is a truism (Eeley, 1955). That unreliability is a major weakness of essay tests in general, and writing samples in particular, also is true. Indeed, Coffman (1971) has suggested that "...a significant portion of the variance of scores on essay examinations consists of systematic or variable error unrelated to the underlying ability one wishes to assess." As early as 1890, the measurement error of essay tests was demonstrated in an experimental study (Edgeworth, 1890) which revealed a wide variation of percentage grades assigned English composition papers by multiple raters. Since that time, analytic and empirical studies of essay tests have uncovered additional sources of error (Yeasmeen and Baker, 1973), which have been variously classified over the years. Perhaps the most comprehensive classifications of error in composition tests requiring writing samples are those described by Braddock, et al., (1963) and French (1972).

French's description divided error into four categories: 1) student error, i.e., variations in performance over time due to the testing situation; 2) test error, i.e., fluctuations in performance attributable to characteristics of the writing task and topic; 3) scale error, i.e., the severity of judgment of individual readers; and 4) reader disagreement, i.e., lack of consistent standards of judgment among raters.

Braddock's scheme is similar to that presented by French, containing the following classifications of error: 1) the writer variable; 2) the assignment variable; 3) the rater variable; and 4) between-rater variability. Braddock's classification scheme, widely cited in the literature on composition research, will be used as an organizing rubric for the following discussion on sources of unreliability of writing samples.

The writer variable. Simply stated, research on fluctuations in writing ability has indicated that the quality of an individual's writing is likely to vary from situation to situation and from time to time. McColly (1970) has pointed to the test situation itself as an influence on writing quality, noting that an examination ipso facto creates for the normal student "...not the best but the worst possible writing conditions." He added, however, that the testing of ability is a "material," not an ideal, process which should be undertaken with normal distractions and pressures. Empirical evidence suggesting that pressures of an examination setting have negligible effects was, in fact, obtained by Kincaid (1953), who also examined fluctuations over time in the writing performance of college freshmen when the topic is held constant. Kincaid concluded that the day-to-day performance in writing of individuals does vary, especially among better writers. On the basis of studies such as

the one conducted by Kincaid, Braddock, et al., (1963) and Diederich (1974) have recommended that at least two occasions of writing take place when assessing English composition skills with writing samples.

The assignment variable. A writing task can be divided into two major components that potentially can influence a writer's performance: topic and mode of discourse (Braddock, et al., 1963). It is an incontrovertible fact that people unfamiliar with a topic will not be able to write intelligently about it, regardless of their potential writing capability. In addition, there is ample evidence, summarized by Lloyd-Jones (1977), to underscore the value of selecting a topic likely to provoke the writer's interest.

An enlightening perspective on characteristics of the assignment variable was obtained by Veal and Tillman (1971) who examined mode of discourse variation in the evaluation of children's writing. Prompted by the fact that most research efforts in English composition have not controlled for mode of discourse, they developed writing tests for grades one through six consisting of one topic in several different modes of discourse. Criteria common to all modes, as well as criteria which were mode-specific, were employed as a basis for scoring the writing samples. They found the expository mode to differentiate most among the student writers, with the greatest overall difference between the expository prose of second and sixth graders. In the Kincaid study (1953) cited earlier, the writing of poor writers was found to vary significantly by topic. Re-analysis of the topics, however, suggested that the critical dimension distinguishing the topics was the mode of discourse, rather than the content.

Clearly, more research is needed before statements regarding the effects on performance of the assignment variable can be made with certainty. Until such evidence is available, it behooves researchers in composition to at least control for differences in the conceptual integrity and content of writing topics, as well as mode of discourse.

The rater variable. From the literati to the barely literate, reader response to written discourse is largely unpredictable and difficult to prescribe. Readers do vary standards for judging the quality of written language from one piece of writing to the next; readers also are likely to change their opinion of a single piece of writing when they read it for a second time. These phenomena make it difficult to objectively and reliably judge the writing of students for evaluative purposes. Eels (1930), for example, demonstrated that the variability in essay test scores assigned by the same reader on different occasions approached the degree of variability of marks assigned by different readers.

Specifying the criteria against which a piece of writing ought to be judged has proven to be one of the most effective means to reduce the influence of within-rater variability. Further research (Stalnaker, 1934) has indicated that raters must have practice, regardless of the type of scale or checklist used to judge writing samples. Speed and rate of reading also have been identified as variables that can influence a rater's judgment. On the basis of a study in which 1200 essays were judged by sixteen high school English teachers, McColly (1970) concluded that "...the faster the rate and speed the more valid and reliable the judgment."

The readability of a piece of writing, including handwriting, format, and mechanical (e.g., grammar, punctuation) correctness, also have been found to

influence reader judgment (e.g., Remondino, 1959; Klein and Hart, 1968).

In addition to preserving anonymity of the writers, Stanley (1962) and others (e.g., Braddock, et al., 1963; Cooper, 1977) have suggested that writing samples to be evaluated should be arranged in random order if more than one occasion of writing is to be judged; that the order of occasions should be concealed from the raters; and that papers from both occasions should be judged together, after all writing samples have been obtained.

The colleague variable. Braddock's classification of the "colleague variable" refers to the tendency of several raters to vary from each other in their judgments regarding the quality of a piece of writing. Perhaps the most widely cited study revealing the differences of opinion that prevail in uncontrolled grading situations is that conducted by Diederich, et al., (1961). Three hundred expository essays (150 essays on each of two topics) written by college freshmen were judged by fifty-three readers drawn from six occupations: professors of English, social science and natural science; lawyers; people in business; writers and editors. The median correlation of the readers was .31. While judgments of the English professors did reflect more unanimity than the other groups, the median correlation of readers in English was only .41, substantiating McColly's (1970) contention that there is "...dramatic disagreement among the hundreds of thousands of English teachers as to what good writing is and is not." It should be remembered, though, that the raters in the Diederich study were not supplied with a set of criteria against which the essays were to be judged.

One would expect that proper training and orientation of raters, plus the use of clearly defined and well-understood criteria would increase reliability

estimates of scoring procedures. This notion has in fact been demonstrated by Stalnaker (1934) who had an undisclosed number of college English instructors carefully reread a composition examination after a period of training. He found that rater reliability on the first reading was as low as .30 and never as high as .75 but that, after training, the reliabilities on the second reading ranged from a low of .73 to a high of .98 with an average of .88.

Unfortunately, though, even a clearly explicated scale or checklist is not a panacea for problems of unreliability, as demonstrated by a study conducted by Follman and Anderson (1967). In this study, five groups of five raters, education-English majors, were paid to assign grades to ten themes, each about 370 words long, from high school and college students of a wide range of abilities. Five methods of rating were employed: 1) the California Essay Scale, in which twenty-five questions about content, style, and mechanics are asked; 2) the Cleveland Rating Scale, in which content, conventions, and style are rating on ten items, such as "organized versus jumbled;" 3) the Diederich Rating Scale, in which points are given for eight dimensions ranging from ideas and organization to spelling and handwriting; 4) the Follman English Mechanics Guide, which is a checklist for such aspects of writing as punctuation, sentence structure, diction, and usage; 5) the Everyman Scale, in which raters employ their own criteria.

The investigators found that the essays received essentially the same scores from all five rating groups. From this finding, they inferred that the rating scales measure many common elements and that the usual unreliability of writing sample evaluation occurs as a result of the heterogeneity of the experiential and academic backgrounds of the raters.

Guided scoring procedures. Nonetheless, guided procedures for sorting or ranking written pieces do have the advantage of making public the criteria of evaluation. Braddock, et al., (1963) observed that three basic types of procedures commonly are employed, although there is considerable variation in their content and administration.

One type of procedure involves comparing a piece of writing with a carefully selected set of compositions ranging in quality. The difficulty of achieving one-to-one correspondence on elements between the criterion pieces and the piece being evaluated makes this procedure cumbersome for large scale assessments, though it has utility as a means of achieving consensus regarding the desired qualities of a writing assignment.

Another procedure requires raters to assign points to each of several specified aspects of a composition. Cooper (1977) has referred to this type of rating procedure as an "analytic scale" which lists "...the prominent features or characteristics of writing in a particular mode." Analytic scales share common qualities with the "atomistic" scales described by Lloyd-Jones (1977). For example, a complex atomistic scale is one which specifies elements of a composition, such as organization, evidence, or diction, within a particular mode of discourse. Lloyd-Jones has suggested that atomistic scales which require the rater to consider named features of discourse are "...tied to abstract categories associated with traditional rhetoric" and do not necessarily reflect the context of writing, i.e., the audience, purpose, and author's voice.

Analytic scales usually involve two or more independent ratings. Typically an overall rating is obtained from each rater by totalling the points assigned to each element on the scale; the overall rating then is averaged with

the overall rating of the other readers. Various forms of these scales have been developed for use in schools as well as research in composition.

A third type of scoring procedure, described by Braddock (1963) as the "general impression" method and by Follman and Anderson (1967) as the "Everyman" procedure, involves two or more raters independently reading each written piece and applying their own criteria of quality. Scores across raters are then averaged to arrive at a total score for each composition.

A procedure which falls between the precision of analytic rating methods and the ambiguity of impressionistic judgments has been described by Lloyd-Jones (1977) and is referred to by him as "primary trait scoring." Primary trait scoring is not concerned with qualities of writing, such as vocabulary and syntax, that are presumed to be characteristics of all good writing. Rather, the procedure focuses raters' attention on whether a piece of writing has certain characteristics, or primary traits, that are critical to success within a given rhetorical task. Underlying this approach is the notion that excellence in one sample of one mode of writing does not necessarily predict excellence in other modes. A precondition for developing primary trait scoring protocols is to select or generate a model of discourse. Lloyd-Jones, for instance, developed a tri-part model specifying the purpose of explanatory, persuasive and expressive types of writing. The aim of developing or selecting a discourse model is to identify precise characteristics of a writing task which then will serve as the basis for creating "...exercises which simulate writing in the appropriate range [of discourse] but not beyond it." (Lloyd-Jones, 1977) The piece of writing elicited by the exercise is then judged by raters who place a numerical value on the writing. The raters' judgments

are informed by a scoring guide which defines and describes varying levels of quality for the task. Lloyd-Jones' scoring guides require a holistic judgment, i.e., an evaluation of the piece as a whole rather than evaluation of separate elements. Supporting development of the guide is the notion that the scoring criteria should directly reflect demands of a given writing task. The limitation of the approach, however, appears to be an over-emphasis on task-specific features in that the scoring system is developed to match a particular assignment, not a class of assignments within a given mode of discourse. Whether the general paradigm of primary trait scoring can be successfully adapted to fit discourse-specific, not assignment-specific, analytic procedures -- and still maintain a close correspondence between task and criteria -- has yet to be researched.

The Contribution of Indirect Measures of Writing

The preceding section clearly weakens the case of detractors of writing samples who build their case solely with claims concerning the inherent unreliability of writing samples. With careful analysis and review, educators ought to be able to control dimensions of a writing assignment likely to produce variations in performance unrelated to capability. With commitment and time, a relatively high degree of reliability of judgment among raters is likely to ba achieved. Moreover, the alternative to employing a direct measure of English composition skills, e.g., a writing sample, is obviously an indirect measure, i.e., an objective test.

The clear advantage of objective tests of writing is the efficiency with which they can be administered and scored. In addition, it has been proposed that a mixture of writing samples and objective items can provide a powerful information base for making instructional decisions (Snidman and Quellmalz, 1976).

Despite such positive characteristics, objections to the use of indirect measures of writing are widespread. Most of the objections center on the lack of prima facie validity of measurement tasks that do not require examinees to engage in the process being measured. A statement of the position ordinarily taken toward objective tests by scholars and teachers in the field of literature and language was made by Coffman and cited by McColly (1970):

> To reject a test of relatively low reliability which was really a sample of the underlying quality in favor of a more indirect measure of higher reliability might be to settle for lower validity of the measure.

In short, to the extent that direct measures of writing are plagued by potential problems of unreliability, developers of indirect measures are obliged to provide evidence of measurement validity. State-of-the-art strategies used to demonstrate validity can be roughly classified as analytic or empirical. Analytic approaches center on the subject matter within a test, with attention directed toward the content of the items and the skills they purportedly measure. In contrast, empirical strategies typically are aimed at investigating potential relationships between performance on objective items and some external criterion. The following discussion highlights by way of extended examples some of the procedures, and limitations, associated with these two validation strategies.

An analytic strategy: The College English Placement Test. The charges leveled at commercially available objective tests of writing are diverse, but many focus on the fact that objective tests rarely contain items measuring skills beyond the confines of the sentence. Grommon (1976), for example, reviewed three commonly used objective tests of writing (e.g., Tests of Academic Progress, 1964; Basic Skills System Writing Test, 1970; Sequential Tests of

Educational Progress, 1957) and observed that the majority of the items were concerned with sentence structure, word meaning, and writing mechanics.

Efforts to construct items which tap "higher level" English skills have been made by some test developers, notably the College English Placement Test (CEPT) (Haugh and Brown, 1969). The objective CEPT items were structured to parallel what were considered to be the usual steps in writing, beginning with the selection of a suitable subject, followed by "...the problem of organization, the listing of pertinent ideas, the identifying of dominant, subordinate and irrelevant ideas, and the grouping of ideas into systematic form." (Haugh and Brown, 1969). Additionally, the CEPT includes two optional essay assignments which are intended to be scored according to the preferences of the test user.

The power of the CEPT as an indirect measure of writing thus hinges largely on the conceptual framework from which the test items were derived. However, the test developers did not collect evidence to indicate the extent to which the items do in fact approximate the hypothesized process of writing. Moreover, the test developers have not offered documentation concerning the extent to which the content and features of the objective items conform to the initial conceptualization. In fact, the procedures by which items were selected for inclusion in the test are likely to have diminished the conceptual integrity of the measure. The item selection procedure involved determining the difficulty level, i.e., the percentage of students answering an item correctly, with the ultimate aim of selecting a final set of items with an optimum level of difficulty of fifty percent. The application of this procedure resulted in a 106 item test with an average difficulty level of .589. The criteria for

inclusion of an item was, then, the apparent ease or difficulty of items, not their content nor the extent to which they correspond to the test authors' conceptualization of the writing process. Thus, the test developers' efforts to develop a measure that would "...provide insights bearing on the entire process of written composition" (Haugh and Brown, 1969) were diminished by the test construction properties employed as well as the lack of evidence indicating the extent to which the objective items correlate with the "process of written composition."

Empirical approaches: The ETS English Composition Test. Underlying many of the criticisms of objective tests of writing is a conviction that essay examinations make a unique contribution to writing assessment which cannot be yielded by scores from indirect measures. This assumption, that objective items are of limited utility and validity as measures of writing, was challenged by the Educational Testing Service (ETS) in a comprehensive investigation of the measurement of writing skills (Godshalk, 1966). The overall intent of the study was to investigate the relative validity of different objective approaches to the assessment of composition skills, using a criterion consisting of five samples of free writing. A total time of forty minutes each was allowed for two of the criterion essays, while the remaining three were extended paragraphs to be written in twenty minutes. The subject matter of the criterion essays was intended to elicit writing in different forms of discourse -- descriptive, narrative, expository, and argumentative writing. Each sample of writing was read by five different raters who rated the papers holistically, assigning a score of three for a superior paper, two for an average paper, and one for an inferior paper. The twenty

five discrete ratings obtained from the 650 eleventh and twelfth grade Subjects were added to make a total score. Reliability of ratings was calculated at .92.

Eight subtests consisting of objective items also were administered to the Subjects. The subtests covered the following skill areas: paragraph organization; usage; sentence correction; language appropriateness; error recognition; construction shifts; and two subtests requiring the examinees to edit a passage (interlinear exercises). The subtest scores were correlated with the sum of the five scores on each criterion essay and with the total essay score. Reasoning that each type of item was designed to tap different skills believed to be components of writing ability, the investigators also compared essay scores to various combinations of subtests. Thirty-six different combinations were examined. The lowest correlation coefficient produced was .716; for twenty-seven of the thirty-six combinations the coefficient was .740 or better, with .775 as the highest correlation. The investigators concluded that, with the exception of one subtest (paragraph organization), the objective items were excellent predictors of "writing ability."

The findings from the ETS study are impressive and could be construed as convincing evidence for the considerable contribution of objective items to writing assessment. Yet the type of validity demonstrated in the study fails to satisfy the measurement requirements of current competency-based assessment efforts.

For example, one limitation of the ETS study procedures concerns the criterion against which the subtests were validated. As the investigators

stated, the intent of the study was to examine the relative validities of the subtests as predictors of "writing ability,"a construct which was defined by holistic scores derived from five criterion essays. The benefit of holistic scoring is that it permits readers to respond impressionistically, free from any arbitary bias that might be introduced by analytic rating procedures (Page, 1968). On the other hand, holistic scoring, as implemented in the ETS study, precludes any reliable and public disclosure of the aspects of writing taken into account by the raters. Indeed, clearly explicated rating criteria actually constitute an operational definition of the writing task. Without such a definition, it is impossible to determine, with any degree of precision, which writing skills or rhetorical features were the targets of evaluation. In the ETS study, what is meant by the construct of "writing ability" is thus a matter of individual interpretation and is likely to vary greatly from one person to the next.

In addition, the use of holistic scoring procedures precluded the identification and assessment of discrete elements, such as structure, development, or even mechanics, of the writing samples. Consequently, no inferences were -- or could be -- drawn about possible relationships between the classes of skill tapped by the subtests and analogous dimensions of actual writing skill as demonstrated on the writing production tasks. At best, one can conclude that various combinations of skill believed to be components of composition as measured by the subtests are related in some unspecified fashion to actual writing performance. Exactly which components of writing ability the subtests do predict, however, is an uncertainty.

Competency-Based Assessment and the Selection of Writing Tasks

The preceding discussion has suggested that research efforts in composition assessment have not resulted in conclusive findings regarding the relative merits of direct and indirect measurement strategies when the purpose of measurement is to ascertain students' level of competence within a defined domain of writing skill. For example, it is generally acknowledged that writing samples possess prima facie validity; yet, educators also recognize that direct strategies can be costly in terms of administration and scoring time. Findings from empirical studies (e.g., Godshalk, 1966; Breland, 1977) have indicated that indirect strategies can be used to assess some constellation of writing skills; however, there is little evidence to suggest that indirect measures can be used as meaningful estimators of specific skills that are highly valued by professionals in English, such as the ability to organize and develop expository paragraphs.

The need for research on the relative merits of varying writing tasks for competency-based assessment purposes has prompted the proposed study. Procedures of the study will be designed to take into account the purposes commonly ascribed to competency-based assessment programs, such as the California Hart Law. Such programs call for measurement strategies which assess students' mastery of skills within a precisely defined domain of learning and which yield maximum information about specific competencies in minimal testing/scoring time. In keeping with the pragmatic demands of the corrent move toward competency-based achievement assessment, aspects of the measurement strategies manipulated in the study were constrained by the following considerations:

1. Ecological validity. The measurement tasks should represent categories of writing operations typically expected of high school students.

2. Instructional sensitivity. To facilitate understanding of test content and student performance by teachers and other instructional decision makers, attributes of the measurement tasks should be designed to reflect conventionally recognized structural and rhetorical elements of written composition. Aspects of the tasks also should have potential for enhancing the diagnostic value of the measure.

3. Educational importance and need. The measurement tasks should elicit non-trivial behaviors which are regarded as important by teachers and other educational decision-makers and, in addition, should represent high priority instructional targets, i.e., areas of identified skill deficit among high school students.

METHOD

Method -- Overview

The main purpose of this study was to investigate the relative merits of three alternate strategies to measure high school students' expository writing skills. Two of the strategies required secondary students to produce samples of writing. The distinction between the two writing sample strategies lay in the response criteria by which the samples were judged. One form of criteria involved application of an Analytic rating scale which required raters to assign scores to six different characteristics of the writing products. The other form of criteria required raters to apply an Impressionistic rating scale which yielded a single score on the quality of the writing samples as a whole. The third assessment strategy required students to select (rather than produce) a response; that is, the strategy employed what commonly are referred to as objective test items which were derived from the elements included on the Analytic rating scale.

Two parallel-form writing sample tasks and the objective items were administered to all subjects, who were randomly assigned to one of six treatment groups determined by the order in which the measures were administered. Each of the collected writing samples was rated by two pairs of raters employing either the Analytic or Impressionistic rating scales.

The Sample

Subjects participating in the study were 128 students enrolled in six different English classes from three high schools in the Los Angeles area. 104 students were in the eleventh grade (five classrooms) and 24 students

(one classroom) were in the twelfth grade. Three of the eleventh grade classes (N=48) were located in a high school situated in a relatively low socioeconomic status community. The two additional eleventh grade classes (N=56) were obtained from a high school in a middle and upper-middle class neighborhood. A third high school in a predominatly middle-class community supplied the class of twelfth grade students (N=24). Teacher participation in the study was on a voluntary basis. Those who agreed to participate (N=5) were asked to use a five-point scale to rate the writing ability of their class as a whole. Their estimates are presented in Table 1.

Table 1

TEACHER ESTIMATE OF PARTICIPATING CLASSES' WRITING ABILITY

Class	Grade Level	Writing Ability Estimate
I (N=12)	11	3
II (N=24)	12	2
III (N=17)	11	3
IV (N=23)	11	3
V (N=33)	11	1
VI (N=19)	11	5

1=Far above average
2=Slightly above average
3=Average
4=Slightly below average
5=Far below average

As indicated in Table 1, the writing ability of participating classes, as estimated by teachers, spanned the range from far above to far below average.

Treatment Groups

Students were randomly assigned within each class to one of six different treatment groups. The treatment groups were determined by the order in which the three measures (e.g., writing sample on Topic 1, writing sample on Topic 2, and the objective item test) were administered. There were three separate measurement occasions within each class, with two days as the least amount of time elapsing between occasions and three days as the maximum time separating occasions. Figure 2 depicts the six treatment groups employed within each classroom.

Figure 2

TREATMENT GROUPS WITHIN EACH CLASSROOM

Treatment Groups	Measurement Occasions: 1	2	3
1	Topic 1	Topic 2	Objective Items
2	Topic 1	Objective Items	Topic 2
3.	Topic 2	Topic 1	Objective Items
4	Topic 2	Objectives Items	Topic 1
5	Objective Items	Topic 2	Topic 1
6	Objective Items	Topic 1	Topic 2

Study Procedures -- Overview

Prior to data collection, each teacher received a notice explaining the procedures of the study and a schedule indicating dates and times of the measurement occasions. (See Appendix A.) Data collection for all six classes occurred over a two week period of time near the conclusion of the academic school year. Collection of data, including administration of the measures and supervision of the students, was performed by the researcher and

four graduate students from the UCLA Graduate School of Education. One data collector was present in each class for each measurement occasion.

Three separate packets of materials, one packet for each measurement occasion, were assembled for each class. The packets contained a class roster with students' names, code number and treatment group, a script for the data collector, directions concerning administration of the measures, copies of the measures, and extra pens. To ensure that students received the appropriate measure according to treatment group assignment, each copy was labeled with the student's name, code number, treatment group, and the date on which the measure was to be administered. The objective tests were printed on white paper, Topic 1 on yellow paper, and Topic 2 on pink paper to facilitate the ease with which directions were given.

Prior to the first data collection activity, teachers were asked whether they wished to have copies of their students' writing samples and, if so, whether they intended to evaluate the samples as part of the students' course grade. All teachers wished to receive copies of the samples, but none intended to use them for grading purposes. Xerox copies of the collected samples were made immediately after each measurement occasion and were returned to the teachers within three days. Students in all classes were informed that their essays would be read and reviewed by their English teachers.

Procedures for all classes for each of the three measurement occasions were similar, with the exception of Occasion One when the data collector gave a brief introduction to the study. Students were assured that no student, teacher, or school names would be reported. The data collector asked that the class remain quiet until the end of the period and explained that questions would be answered individually. After the measures were distributed

the data collector first gave directions for writing the essays (Topics 1 and 2), followed by directions for completing the objective measure. Students then were requested to begin working. A copy of the data collectors' scripts is included in Appendix B.

The remainder of the class period was devoted to completing the measurement tasks, with an announcement of the time by the data collector when ten minutes were left. The amount of time allowed ranged from thirty-five to forty-five minutes according to variations in the schools' schedules. The average amount of time was approximately forty minutes.

Measures

The Writing Task

The writing task employed in this study can be divided into two major components: the writing stimulus (i.e., topic and directions) and the response criteria by which writing samples were judged. In this study, two writing topics were developed and two sets of response criteria, Analytic and Impressionistic, were applied. Each of the writing task components are discussed in separate sections below.

The Writing Stimulus (Topics and Directions)

Two topics, designed to elicit like-samples of student writing, were developed. The topics were intended to promote writing within the discourse domain of exposition, that is, writing "...that explains or clarifies a subject" (Brooks and Warren, 1961). Task attributes constraining the development of the topics included discourse mode, rhetorical purpose, content limits, and intended audience. Appendix C presents the attribute specifications which guided the development process.

Both topics presented students with a one paragraph description of a problem situation. The situation presented in Topic 1 concerned the decline in the basic skills of reading, writing, and arithmetic among high school students, while Topic 2 presented the problem of the abuse of alcohol, drugs, and cigarettes by young people. Both Topics then instructed students to assume that they had been asked by a group of parents and teachers to explain the problem. Topic 1 instructed students to "explain why many of today's students graduate from high school without basic skills." Similarly, Topic 2 instructed students to "explain why many of today's teenagers abuse alcohol and drugs or smoke cigarettes." In addition, both Topics urged students to "back-up...ideas with support, such as examples, facts, and other details." Appendix D presents copies of both Topic 1 and Topic 2.

Development of the topics was informed by the results of several cycles of pilot testing. A total of eight different topics (four sets of parallel topics) were developed. Four separate pilot tests were conducted, each employing one parallel set of topics, in four tenth and eleventh grade English classes ranging from thirty to fifty students per class. The two topics were randomly assigned to students within each class on each pilot test occasion. Copies of the pilot test topics are presented in Appendix E.

The main purposes of the pilot tests were to determine: 1) whether the content called for by the topic was appropriate to the students' background and experience; 2) whether the directions to students were clear and easily understood; 3) whether the topic elicited the desired class of responses, e.g., expository writing. Specific questions addressed after each pilot test were as follows:

1. Does the writing task elicit writing in an expository mode, rather than description or narration?

2. Does the task encourage a class of responses that lend themselves to development with evidence or other details? That can be logically developed?

3. Are directions to the students clear? Should students be given more information about what is expected of them? Less information?

4. Are the topics prompting students to write essays of sufficient length?

5. Are students able to write a complete essay within the time allotted?

6. Do students have sufficient experience or background information to write a reasonably well-informed essay? Or, does the topic appear to elicit only opinions?

The initial pilot test (Topics A and B) required students to write only one paragraph. Review of the collected samples indicated that students tended to write only four or five sentences. Paragraphs of this length did not provide a sufficient sample of writing for judging organization and development of ideas within the writing sample.

Topics included in the second pilot test (Topics C and D) were similar to those employed during the first test, with the additional requirement that students write at least one page. While most students were able to write approximately two hundred words within thirty-five to forty minutes, review of the samples revealed that students tended to respond to the topic with only statements of opinion. Few students attempted to provide reasons, examples, or other support to explain why a particular opinion, or point of view, was held. Thus, the topics were deemed inappropriate because they did not encourage the desired class of responses, e.g., logical development and support for generalizations and opinions.

Topics tested on the third occasion (Topics E and F) were formulated with the aims of 1) calling upon students to take a position, i.e., assume

a point of view, on a subject and 2) requiring them to support their position with specifics beyond personal opinion. Analysis of the collected essays indicated the topics erred too far in the direction of abstraction and distance from personal experience. Apparently, Topic F required specialized information (e.g., consumer rights) not available to many of the students tested, while Topic E encouraged students to provide generalizations and little or no specific support.

The fourth pilot test (Topics G and H) yielded more encouraging results. Review of the collected papers indicated that the students were able to write essays which:

1. attempted to explain or inform
2. were developed through a pattern consisting of generalizations and related support
3. did not rely exclusively on personal opinion for support
4. were written within the allotted time period of forty minutes.

These results may possibly be accounted for by the following writing task characteristics that were not included in earlier drafts of the topics:

1. Students were provided with a clear statement of the essay purpose (e.g., "Remember, the purpose of your essay is to give an informative explanation.")

2. The task stimulus provided a context and audience for the writing (e.g., "Assume you have been asked by a group of parents and teachers to explain the problem...").

3. The content of the topic did not require specialized knowledge and was familiar to most of the students.

Topics G and H consequently were employed as the writing stimuli in the final study. Note that Topics G and H are labelled as Topics 1 and 2,

respectively, throughout this report.

The Response Criteria

Two sets of response criteria (and corresponding rating procedures) were developed, an Analytic rating scale and a guided procedure employing an Impressionistic rating scale. As the name implies, analytic rating procedures "...rely on the assessment of particular features associated with skill in discoursing" (Lloyd-Jones, 1977). Employed in this study was an Analytic scale which provided a separate rating for each of six prespecified elements of writing. In contrast, impressionistic procedures yield a judgment based on the total writing product, with the whole piece presumed to be greater than the sum of its parts. The Impressionistic scale developed for this study thus was designed to provide a single numerical value for each writing sample. Both the Analytic and Impressionistic scales are discussed below in separate sections.

The Analytic Scale. An Analytic scale requiring raters to assign a score of one through four to each of six specified elements of a writing product was developed. The scale thus yielded an estimate of the overall quality of an essay (i.e., the sum of scores assigned to each element) as well as sub-scores on each of the six elements.

Analytic scoring procedures presume that a piece of writing can be viewed as consisting of component, but not necessarily independent, parts which are worthy of individual scrutiny. Depending upon the intended function of the scale, different approaches can be taken to identify the components including, for example, the use of linguistic structures or the results of behavioral analyses of the writing process.

A manor concern of this study was to explore the relative merits of

analytic rating procedures within the current context of competency measurement. In this context, one function of scale scores is to yield information on specific strengths and weaknesses of a writing product. In addition, the scores ought to convey information that is easily interpreted and meaningful to instructional decision-makers. Given these twin functions, an Analytic scale intended to reflect state-of-the-art pedagogical precepts and practice in composition was developed. Elements on the scale were derived from an analysis of conventionally recognized structural features of exposition as indicated in curriculum guides and textbooks at the secondary level.

The final version of the Analytic rating scale was designed to measure the extent to which the collected writing samples met the following criteria:

1. Essay Focus: The introduction or conclusion of the essay clearly indicates the subject and main idea of the whole essay.

2. Essay Development: All major subtopics ("main points") clearly relate to the main idea of the whole essay.

3. Essay Organization: The main idea is developed according to a clearly discernible method of organization.

4. Paragraphing: The essay is composed of one or more clearly discernible units of thought, i.e., paragraphs.

5. Support: Generalizations and assertions are supported by specific, clear, supporting statements.

6. Mechanics: The essay is free of intrusive mechanical errors.

These criteria were amplified and transformed into six separate subscales, with a range of four (high) to one (low) points per subscale. Descriptions of essay characteristics for each of the four levels for all six subscales were developed. An abbreviated version of the Analytic scale is presented below. Please refer to Appendix F for an unabridged version of the scale.

Figure 3

ABBREVIATED VERSION OF THE ANALYTIC RATING SCALE

Element 1: Essay Focus

The introduction or conclusion of the essay clearly indicates the subject and main idea of the whole essay.

4 = The introduction (and/or conclusion) of this paper clearly conveys the main idea of the whole essay. It also limits the topic by alerting the reader to the key points covered in the body of the essay.

3 = The introduction (and/or conclusion) of this paper conveys the main idea of the whole essay. It sets limits on the topic, but does not clearly suggest how the main idea is developed.

2 = The introduction (and/or conclusion) of this paper gives the reader a fairly clear sense of the main idea of the whole essay. However, neither the introduction nor the conclusion help focus -- or bring direction to -- the body of the paper.

1 = Neither the introduction nor the conclusion is helpful to the reader in obtaining any sense of the main idea of the essay AND/OR the main idea of the whole essay is not clearly stated or implied, or no reference is made to the main idea, or the reference is confusing.

Element 2: Essay Development

All major subtopics ("main points") clearly relate to the main idea of the whole essay.

4 = All ideas and details in this paper are clearly and logically related to a single dominant main idea. Development of the idea is not interrupted by digressions of thought or extraneous material.

3 = All major ideas in this paper are related to the dominant main idea. There may be a few minor digressions of thought, such as extended listings of supporting detail. However, digressions do not stray entirely away from the main point; they simply interrupt the logical flow of thought. One or two tangential ideas may be introduced but are not developed.

2 = The relationship of many ideas and details to the overall main idea of the essay is somewhat tenuous. This may be due to one (or more) of the following types of problems. The paper may begin to develop a main idea, but stray or wander off the topic, failing to return to the main point. Or, the paper may be peppered with many tangential ideas/details which disrupt the development of subtopics (and, in turn, the main idea).

Figure 3 (continued)

1 = Few ideas or details in this paper contribute to development of a single main idea. In fact, the reader is likely to have difficulty discerning the main idea of the essay, if one is present. A paper which does not address the topic posed in the assignment also should receive a "1" rating.

Element 3: Essay Organization

The main idea is developed according to a clearly discernible method of organization.

4 = The plan by which the essay is structured is readily apparent and consistently applied throughout the essay. Some logical scheme serves as the underlying basis for the overall essay structure. That is, there is an easily discernible logical flow from one idea (subtopic) to the next.

3 = The essay is organized according to some recognizable plan which is employed throughout the essay. Some organizing rubric is used to divide the content/ideas of the essay into units of thought. For example, each subtopic might represent a different reason or category of content. Subtopics, however, are not presented in any particular logical order.

2 = The essay is not structured according to any readily discernible plan. However, the reader is able to infer some association between major ideas. The logical flow from one idea to the next is so tenuous that the reader must supply his/her own links or transitions to form a clear relationship among ideas.

1 = No plan for organizing ideas and details is apparent. Subtopics are not arranged in any discernible order. The reader has difficulty inferring any relationship among ideas or any pattern of thought.

Element 4: Paragraphing

The essay is composed of one or more clearly discernible units of thought, i.e., paragraphs.

4 = This paper contains one or more clearly circumscribed subtopics which are developed by cohesive groups of supporting statements. Each subtopic represents a complete unit of thought. Major units of thought are delineated by conventions of paragraphs.

3 = This paper is composed of one or more subtopics which are developed by supporting statements. Each subtopic represents a complete unit of thought. Distinctions between major units of thought would be clearer to the reader if paragraphs were more effectively employed.

2 = This paper contains one or more subtopics which represent major units of thought. However, the distinction between different units of thought

Figure 3 (continued)

is not consistently clear. Distinctions may be blurred because topic sentences and related support are not presented as cohesive and logically related groups of sentences.

1 = The essay does not appear to contain any clearly differentiated units of thought, e.g., subtopics and related support. Statements appear to be presented in random order, with no clear relationship between topic sentences and supporting statements.

Element 5: Support

Generalizations and assertions are supported by specific, clear supporting statements.

4 = All generalizations and assertions in this paper have logically related support. Supporting statements clearly are at a greater level of specificity than the generalizations they are intended to support. The writer makes effective use of concrete detail to support ideas.

3 = Most generalizations and assertions in this paper are supported by specific, clear and logically related statements.

2 = Support for many generalizations and assertions in this paper is not especially convincing. Supporting statements lack precision, either in language or in the use of specific detail.

1 = Support is provided for few or none of the generalizations or assertions. Few (or none) of the supporting statements are at a greater level of specificity than the generalizations themselves.

Element 6: Mechanics

The essay is free of intrusive mechanical errors.

4 = The writer appears to have control over the mechanical aspects of this essay.

3 = Mechanics are not a problem in this paper. However, the writer appears to be occasionally careless and makes a few minor errors which are readily evident to the discerning reader. The paper also may contain a few common errors in mechanics.

2 = This essay is flawed by errors in mechanics, although the paper does not strike the discerning reader as being illiterate.

1 = Mechanical errors make this paper very difficult to read and understand.

Development of the Analytic rating scale proceeded in two major phases of activity. During the first phase, a review of literature on exposition, elements to be included on the scale were identified. The second phase involved review, application and refinement of draft versions of the scale.

During the literature review process, two characteristics, referred to as "unity" and "coherence," consistently emerged as basic features of expository writing. According to Warriner and Griffith (1973), unity is "...achieved by discussing only one topic" in an essay. Similarly, Gallo and Rink (1968) suggested that unity is promoted by relating supporting statements to the main idea of an essay. The main idea thus can be viewed as the dominating concept, or point of view, and as a generalization which is supported or amplified in the body of an essay. Gallo and Rink (1968) further indicated that unity is enhanced when an essay contains a "controlling idea" which focuses or narrows the subject.

A second dimension of exposition identified during the review process was termed "coherence." Coherence refers to the logical relationship among ideas and can be achieved by arranging supporting sentences and details in some appropriate order. For example, among orders identified by Warriner and Griffith (1973) were chronological, spatial, order or importance, comparison and contrast. The coherence of an essay also is enhanced by the use of linking expressions and connectives (e.g., Gorrel, 1965; Warriner and Griffith, 1973; West, 1973). The literature review resulted in the six criteria described earlier.

The criteria, in turn, served as the basis from which the following three draft versions of the Analytic scale were developed: 1) a dichotomous scale which directed raters to indicate whether specified features were

or were not present; 2) a scale with a range of one through six points for each subscale; 3) a scale with a range of four points for each subscale.

As draft versions of the scale were developed, they were reviewed by members of the research staff of the Center for the Study of Evaluation as well as composition teachers at the high school and post-secondary levels. Reviewers were asked to consider the extent to which the rating criteria represent valid, non-trivial standards for judging exposition at the high school level. Elements on the scale drafts were judged to be valid components of exposition and worthy of evaluation. No additional elements were suggested nor were recommendations to delete elements made. Comments on the metrics of the drafts resulted in the most significant revisions. The dichotomous scale was considered to lack sensitivity to the range of quality that student essays were likely to exhibit. A six point scale for each element then was developed, with LOW = 1-2 points, MIDDLE = 3-4 points, HIGH = 5-6 points. This scale proved difficult to apply to actual samples of student writing, as the distinctions between 1-2, 3-4, 5-6 were not clearly distinguishable. A four-point scale, with descriptions of essay characteristics accompanying each scale point for six different elements, then was developed.

This draft, a prototype of the final Analytic scale, was applied to samples of writing and several cycles of application-revision occurred. The focus of most revision efforts lay in the definition of levels of quality within the six subscales. Final modifications in the scale occurred during the rater training session. During the first phase of rater training, the two raters read, rated, and discussed the ratings of approximately fifty student papers. On the basis of their comments, a condensed version of the scale was developed

which highlighted critical distinctions between levels of quality within subscales as well as differences between subscales. A copy of the condensed version of the Analytic scale is presented in Appendix G.

The Impressionistic Scale. The second form of response criteria was an Impressionistic rating scale which required raters to make a judgment regarding the overall quality of each writing sample. Such an approach "...assumes that each factor that makes up writing skill is related to all other factors and that one factor cannot easily be separated from others." (Office of the Los Angeles County Superintendent of Schools, 1977)

It should be recalled that a pivotal concern of this study involved a comparison of writing sample scores obtained through impressionistic rating procedures and those yielded through application of an analytic scoring system. Consequently, it was critical for the impressionistic rating procedures to be developed independently and without reference to any of the attributes specified on the Analytic rating scale. To minimize possible confounds with Analytic ratings of the writing samples, an Impressionistic Scale containing a minimal rubric was developed.

The discourse domain of concern to this study was expository writing. The rubric of the Impressionistic Scale thus required raters to consider the following question when rating each writing sample: "To what extent is the essay an example of effective exposition?" The rubric also directed raters to assign each essay a single numerical score by employing the following six point scale:

High

6 = An excellent example of exposition

5 = A good example of exposition

Middle

4 = An adequate example of exposition

3 = A minimally adequate example of exposition

Low

2 = A poor example of exposition

1 = Barely readable and/or off the topic

In addition to the scale, the rubric contained several definitions outlining prominent conventionally recognized features of exposition. It was expected that these definitions would serve raters as a common frame of reference for determining what was meant by "exposition."

A try-out of the Impressionistic scale was conducted with the aim of determining whether the rubric and scale provided enough guidance to allow raters to achieve consensus in ratings. Thirty student essays written on the two study topics were rated and discussed by two raters. Through this process, it was determined that the rubric (including the definitions) supplied a sufficient frame of reference for assigning the essays to the three major scale categories of High, Middle, and Low. Discussion and analysis of specific features of individual papers proved to be the most fruitful means for achieving consensus on numerical scores within the categories. These conclusions suggested that at least fifty different essays should be employed as practice papers during actual training of raters and that the training session should allow sufficient time for raters to discuss the essays at regular intervals.

The Objective Item Measure

A thirty-seven item multiple-choice test was developed to measure skills presumably related to the production of written discourse. The skill areas included on the test were intended to be analogous to the criteria specified on the Analytic rating scale. The intent of the test items was to measure, at the level of recognition and discrimination, those categories of skill assessed at the level of production, i.e., actual writing, by the Analytic scale.

The skills to be assessed by means of the objective measures were identified through two related analyses. The first analysis consisted of a review of expository writing skills frequently emphasized in secondary composition curricula and instructional materials and for which selected-response type practice was provided. For example, a typical exercise required students to identify from a list those details which either do or do not support a given generalization. The skills identified through the review were then arrayed against the elements listed in the Analytic rating scale in order to determine which skills were conceptually analogous to the rating scale elements. A final listing of five skills areas (corresponding to the first five elements on the Analytic scale) was compiled, and specifications for item generation were developed.

The final version of the objective measure was comprised of five subtests corresponding to the following Analytic scale elements: focus, development, organization, support, and paragraphing. A subtest analogous to the mechanics element was not developed as mechanics represent a cluster of diverse discrete skills, e.g., punctuation, grammar, diction. To include an adequate sample of test items for the array of skills subsumed under mechanics would result in a lengthy battery of subtests beyond the scope of this study and beyond the available testing time.

Time constraints imposed an additional limitation on the development of analogous test items. The Analytic rating scale assessed writing skills at the level of the essay or long paragraph (approximately 200 words). Ideally, then, the stimulus passages included on the objective measure should be equivalent in length and structure. In order to provide a variety of stimulus passages within the allotted testing time, however, stimulus

components of the test items were designed to reflect the structure of seventy-five to one hundred word paragraphs. The relationship between the objective items and the Analytic scale elements is displayed in Figure 4.

(Insert Figure 4 here)

As Figure 4 suggests, the categories of tasks included on the objective measure represent recognition and discrimination skills which presumably are precursers of analogous categories of production skills.

The generation of test items was guided by a set of specifications, included in Appendix H, outlining the stimulus and response attributes (including distractors) of items, as well as item format. To minimize the possible effect of reading ability, items and directions were written at a level that could be easily understood by low ability tenth grade students. A copy of the objective measure is presented in Appendix I.

The final version of the measure contained six separate sections. Each section included items of similar format which were preceded by a set of directions. To reduce the potential effects of order on performance, copies of the tests were assembled with sections presented in six different orders, as shown in Figure 5.

Figure 5

FORMS OF THE OBJECTIVE MEASURE

Orders:	1st	2nd	3rd	4th	5th	6th
Sections	1	2	3	4	5	6
	2	3	4	5	6	1
	3	4	5	6	1	2
	4	5	6	1	2	3
	5	6	1	2	3	4
	6	1	2	3	4	5

Figure 4

RELATIONSHIP BETWEEN ANALYTIC SCALE ELEMENTS AND OBJECTIVE TEST ITEMS

ANALYTIC SCALE		OBJECTIVE ITEMS
I. The introduction or conclusion conveys the main idea.	Can the student infer the main idea of a given passage?	I. Given a brief passage, student identifies the main idea. (Items 1-5, 26, 30, 34)
II. Major subtopics relate to the main idea.	Can the student identify statements that are not logically related to topic sentences?	II. Given a topic sentence, student identifies details which do not develop or explain the topic sentence. (Items 11-15, 27, 31, 35)
III. The main idea is developed according to a logical method of organization.	Can the student arrange statements into a logical order?	III. a) Given a series of topic sentences and related support, the student can arrange them in a logical order. (Items 20-25) b) Given a passage with faulty organization, the student can identify the statement(s) which are out of order. (Items 28, 32, 36)
IV. The essay contains one or more units of thought, i.e., paragraphs.	Can the student identify where in a passage a complete thought ends and a new one begins?	IV. Given a passage, the student can identify where a paragraph could begin. (Items 16-20)
V. Generalizations are supported by specific support.	Can the student identify statements which best support a generalization?	V. a) Given a topic sentence, the student can identify the best supporting statement. (Items 6-10) b) Given a passage, the student can identify which statements require additional support. (Items 29, 33, 37)

Part One (five items) presented students with a 75 to 100 word paragraph lacking an explicit statement of the main idea, followed by a series of four statements. The correct response was a statement at a greater level of generality than all statements included in the stimulus passage and which was directly related to the subject of the paragraph. Distractors were statements relating to one or more (but not all) sentences in the stimulus paragraph.

Part Two (five items) required students to select the one statement from a list of four which best supported a given generalization. The correct response was a statement clearly subordinate to the generalization which directly explained or illustrated the main point. Incorrect statements were related to the subject of the given generalization but provided tangential information and/or failed to explain or illustrate the main point.

The third part (five items) presented a topic sentence followed by four statements. Three of the four statements were directly related to, and helped support, the given topic sentence. Students were to identify the one statement which did not clearly explain, develop, or illustrate the topic sentence.

The stimulus components of items in Part Four (five items) provided students with four to five sentence passage related to one overall main idea. However, within each passage, two distinct thoughts were expressed. Students were directed to identify where "one complete thought ends and a new one starts" (i.e., where a new paragraph could begin).

In Part Five (five items) students were presented with sets of five statements which, according to the directions, represented "lists of ideas and details" that might be included in different essays. The directions instructed students to indicate in which order the statements could be pre-

sented to make a logical essay. In the correct order for each set, the main idea statement was to be placed first or last, and related subordinate statements were to be grouped together as topic sentence and supporting statements. The scoring key for each item provided for arrangement of the statements in either inductive or deductive orders.

The method by which items for Part Six were generated, as well as the item format, was different from that of the first five parts. Unlike the stimulus passages of the first five parts, which were generated specifically for the objective test, the passages in Part Six were drawn from actual samples of high school students' writing. The passages selected exhibited one or more of the following error categories:

1. The main idea was not stated or implied. (Focus)
2. The essay contained at least one digression of thought from the main idea. (Development)
3. There was at least one statement which did not follow the logical structure of the whole essay. (Organization)
4. The essay contained at least one generalization that was not developed with supporting statements. (Support)

Each passage was followed by four items directing the student to identify the statement(s) which exhibited a specified category of error, as outlined above. Note that each error category corresponds to a skill area covered in the first five sections of the objective measure as well as to the first five elements on the Analytic rating scale.

Development of the objective test proceeded through two major phases of activity. The first, or preliminary, phase involved review by staff members of prototype items and pilot testing with small groups of students. Revisions and modifications in the instrument were made on the basis of in-

formation gathered during these preliminary activities. Pilot testing of the revised instrument in several secondary classrooms and review of findings constituted the second phase of instrument development. The main activities for each phase and their implications for refinement of the instrument are summarized below.

The focus of activity during the preliminary phase was to develop a prototypic set of stimulus and response attributes for test items to measure the skill areas below:

1. Identifying the main idea of a given paragraph.
2. Selecting statements to support a given main idea.
3. Identifying statements which do not develop/explain a given main idea.
4. Differentiating between main idea statements and statements of supporting detail.
5. Selecting statements that provide the most concrete and specific support for a given generalization.
6. Identifying where in a given passage a complete thought ends and a new one commences.
7. Selecting the most logical order for a series of given sentences.

A review by CSE staff members indicated the reading level and structural complexity of the stimulus passages was appropriate for secondary students of varying abilities (i.e., not too difficult).

The first draft was administered separately to two small groups (six students each) of average ability eleventh grade students under supervision of the researcher. Students were instructed to stop after completing each section of the instrument to allow for discussion of difficulties and confusion which they encountered. Below is a summary of information and decisions derived from this procedure.

- The reading level was not too difficult.
- Distractors (i.e., wrong answer options) for several of the items were judged to be too obvious. As a consequence, it was decided to hold constant the subject of distractors for each item and to vary only the object or action (i.e., the "main idea") of distractor statements.
- Students found the format confusing for items requiring discrimination between "main idea" and "supporting detail" statements. An alternate format was developed.
- Students were unfamiliar with the term "concrete and specific" support. This item-type was eliminated and replaced with items requiring students to select the "best" supporting detail from among several alternatives.
- The directions for selecting the most logical order for a series of sentences proved confusing. The response format was altered to allow students to mark an order for each individual sentence rather than selecting one order from among several given options.

A revised version of the instrument was assembled and pilot tested in a low ability level eleventh grade class (N=10) and an above average ability level tenth grade class (N=30). Minor modifications were incorporated into the instrument on the basis of these pilot sessions, and the instrument was again administered to an average ability 11th grade class. At the conclusion of each session, comments regarding difficulties encountered were solicited from students. The main purposes of these administrations were to 1) determine whether directions and item format were clear; 2) determine whether the reading level was appropriate; 3) obtain information on the pattern of student

responses (e.g., are some distractors consistently selected? Is there systematic variation in group performance according to ability level?) A summary of findings from the pilot sessions is presented below.

- With one exception, item format and directions proved to be clear. The exception, items requiring students to select a logical order for a series of given sentences, was re-formated and the directions revised.
- No student reported difficulty in reading the stimulus passages and response options.
- All students were able to complete the test within a thirty-five minute period of time.
- Response patterns, which were analyzed at the class level, showed very little variation between low and average ability students.

This last finding prompted a critical review of test content. One possible explanation to account for the minimal amount of variation appeared to lie in the level of complexity of the stimulus passages. In an effort to minimize potential effects of reading ability on test performance, the passages had been constructed according to the following constraints: 1) a limit of four sentences per passage; 2) simple sentence structure (i.e., few embedded clauses, subordination, coordination); 3) simple paragraph organization (e.g., either part/whole development or generalizations followed by a series of related specifics). It was determined that perhaps these constraints were imposing an artificial ceiling on the instrument. The instrument consequently has been revised to systematically increase the complexity of stimulus passages by: 1) increasing length of passages; 2) providing more variation in sentence structure; 3) modifying and adding passages

with alternative organizational structures (e.g., comparison/contrast; development through reason and example).

Rating Procedures

As described under the section, "Study Procedures," the design of this study called for each student to write two essays, one on Topic 1 and one on Topic 2. A total of 117 essays were collected. Students who wrote only one essay were dropped as cases for analysis of the writing samples. This resulted in a total of 112 cases for analysis (224 essays). The completed essays ranged in length from one half of a double-spaced typewritten page (approximately 100 words) to three double-spaced typewritten pages (approximately 600 words). The average length was approximately one and one-half typewritten, double-spaced pages, or 250 words.

To conceal student, class, and school identity, each student's Topic 1 essay was assigned a code number using a table of random numbers. The same random assignment process was followed for coding each student's Topic 2 essay.

All essays were typed, double-spaced, preserving all student errors. All typed essays were proof-read to ensure that the typed version was an exact replicate of the original handwritten essay. The essay code number, followed by a 1 or 2 to designate topic, was typed in the upper left-hand corner of each paper. Three sets of Xerox copies of all typed essays were assembled. Papers within the two sets used by the raters were randomly shuffled so that the essays would not be read in the same order during the rating sessions. Thirty common essays from each set of 224 papers were assembled to allow raters to read the same essay at periodic intervals during

the rating sessions.

Impressionistic Scale Rater Training

Two high school English teachers with previous experience in systematic evaluation of student essays were selected as raters. Both raters were remunerated with stipends for the time spent in training and in the rating session.

The duration of the training session was approximately four hours. Each rater was supplied with a packet of materials containing an introduction to impressionistic rating procedures, four brief definitions of exposition, copies of the two topics, the Impressionistic scale, and approximately fifty typed high school student essays on Topic 1 and Topic 2 to serve as practice essays. In addition, raters were informed that "...the papers may not present a 'normal' distribution, that is, some [scale] categories may have very few papers while others may have many." See Appendix J for copies of the Impressionistic rater training materials.

The raters were informed that one of the primary objectives of the training session was to achieve a high degree of agreement on scores for the independently rated papers. As previously described, the Impressionistic rating rubric provided minimal criteria for assigning scores. Consequently, procedures of the training session were designed to allow raters to inductively arrive at a common frame of reference for assigning scores. The basic format of the training session required raters to independently read and rate sets of practice essays, to compare ratings, and to discuss discrepancies in the assigned scores.

The raters first read one set of ten essays, sorting them into three groups, Low, Middle, and High. After sorting, discrepancies in ratings were

discussed. The same procedure was followed with a second set of ten essays. The raters then returned to the initial packet of ten essays in order to independently assign each essay a numerical value within each of the three categories (e.g., 1-2, 3-4, 5-6). Discrepancies were then noted and discussed, and the rating procedure was repeated with the second packet of ten essays. An additional set of papers then was read and assigned scores of one through six.

Analysis of specific features of individual papers proved to be the most effective means for achieving consensus on numerical scores within the three categories. It was during such exchanges that the raters verbalized their criteria of judgment as well as the relative weighting they placed on particular features of discourse. Issues raised, for example, included the weighting of a creative response versus straying from the assigned topic; errors in grammar and diction in relation to a logically developed essay structure; extraneous ideas and details combined with convincing support.

Toward the conclusion of the training session, the researcher asked the raters to provide brief discriptors for each of the six levels of quality on the scale. These descriptions, displayed in Figure 6, then served as a guide for the actual rating session.

Figure 6

DESCRIPTORS FOR SIX POINT IMPRESSIONISTIC RATING SCALE

6 = The essay excells in all aspects.

5 = The essay is more than adequate and may even excell in some aspects (but not all). A "solid" paper.

4 = The essay has no major deficiencies but also no outstanding qualities. Likely to be "thin."

Figure 6 (continued)

3 = The essay demonstrates marginal competence in most areas. Likely to be uneven in quality. The essay has some problems.

2 = The essay is clearly deficient in many areas.

1 = The essay contains gross, distracting errors in almost all aspects. And/or the essay is off the topic.

At the conclusion of the four hour training session, a pilot test to determine the degree of rater agreement was conducted. Raters were presented with identical packets containing copies of Topic 1 and Topic 2 essays written by fourteen high school students. Thus, each rater read one pair of essays for each student. The essays were typewritten and coded in the manner described under "Rating Procedures." Approximately one hour was devoted to reading and rating the pilot test essays.

Coefficients of rater reliability were calculated for the Impressionistic rating procedure. The method for calculating the coefficients was derived from Cronbach's theory (Cronbach, et al., 1972) for estimating the generalizability of observations. This theory requires identifying sources of variance in observed scores and estimating the magnitude of their individual and combined effects. Sources of variation, or "facets," are conditions in a design which are likely to influence observations. For purposes of analysis, all training and final study rating procedures employed in this study were treated as independent, completely crossed designs, as depicted in Figure 7.

(Insert Figure 7 about here)

The two main facets of concern in the design shown in Figure 7 were form (F) and rater (R), with students and raters assumed to be random effects. The two writing topics (Forms 1 and 2) were viewed as defining the performance

Figure 7

DESIGN OF IMPRESSIONISTIC AND ANALYTIC RATING PROCEDURES

	Impressionistic		Analytic	
Students	Raters: 1	2	3	4
	Forms: 1,2	1,2	1,2	1,2
S^1				
S^2				
S^3				
S^n				

domain of concern to this study, hence, forms were assumed to be a fixed effect. Under this mixed effects model, reliability of ratings was derived as follows:

$$\text{Reliability of raters} = \frac{\sigma^2 s}{\sigma^2 s + \frac{\sigma^2 sr}{nr}}$$

where $\sigma^2 s$ is the variance component of student scores and $\sigma^2 sr$ is the variance component due to the interaction of student and rater. Appendix K includes tables displaying estimated variance components for S, F, R and their interactions. Rater reliability based on these variance components was calculated to be .96. This was deemed to be a sufficiently high reliability estimate. The final study rating session thus was scheduled for the following day.

Impressionistic Scale Rating Session

Impressionistic rating of the 224 student essays took place during a single session of approximately eight hours. Each rater was provided with a complete set of the typed essays as described in the section, "Rating Procedures." To minimize reader fatigue, breaks were scheduled at one hour intervals. Raters were instructed to read and rate the papers in the given order and to not discuss the essays prior to rating.

As the essays were rated, the researcher maintained a tally of the ratings on a master sheet. When discrepancies of more than two points were observed the rating was interrupted in order to discuss the essays. No score was changed as a result of the discussions. Approximately every half hour, raters were requested to read and rate several common essays. After the scores were recorded, features of the essays were briefly discussed with reference to the rating scale.

Rater reliability, calculated for the 224 essays using the method de-

scribed in "Impressionistic Scale Rater Training," was .87, slightly lower than the pilot test reliability. This difference may be attributed to the fact that the final study papers exhibited a greater range of quality and characteristics than the twenty-eight pilot test papers. Variance components for the final study Impressionistic ratings are presented in Appendix L.

Basic Measurement Properties of Impressionistic Scale

A score for each case was obtained by summing the scores assigned to each topic by each rater, creating a range of possible points of 4 (low) to 24 (high). Means and standard deviations of the total population and by class are presented in Table 2.

(Insert Table 2 about here)

The means of Class II ($\overline{X}$ = 15.6) and Class V ($\overline{X}$ = 16.6) were above the population mean of 13.3. These two classes were rated by their respective teachers as above average in writing ability. (See Table 1.) The means of Class IV ($\overline{X}$ = 12.4) and Class VI ($\overline{X}$ = 10.6) were below the population mean, which also was consistent with the teachers' ratings. Both Classes I and III were rated by their teachers as "average," yet their mean performances were below the population mean.

Analytic Scale Rater Training

Analytic scale raters were two teachers of English composition who had previous experience in the systematic assessment of student essays. Each rater was remunerated with a stipend for the time devoted to rater training and to the rating of essays.

The objective of the approximately five hour training session was to promote rater agreement in the application of the six element Analytic rating scale. Rater training materials included an abbreviated version of the Analytic rating rubric, a copy of the entire rubric, a copy of both Topics

Table 2

MEANS AND STANDARD DEVIATIONS ON IMPRESSIONISTIC SCALE SCORES

CLASS:	I	II	III	IV	V	VI	TOTAL
Mean	8.9	15.6	11.3	12.4	16.6	10.6	13.3
SD	1.3	2.7	1.7	3.7	2.5	2.5	3.7
N	10	19	16	19	32	16	112

1 and 2, and approximately fifty typed high school student essays on Topics 1 and 2. As in the Impressionistic rater training, raters were informed that scores assigned to the essays might not represent a normal distribution.

Plans for the training session required raters to incrementally apply the rating scale to the practice essays. After briefly reviewing the six elements described in the rubric, the raters rated several sets of essays on the first element, "essay focus." When a common understanding of the first element was achieved, the raters then applied the rubric of the second element, "essay development," to additional sets of practice essays. After agreement on the application of "development" was reached, the raters then applied both the "focus" and "development" rubrics to approximately five papers, followed by discussion of the ratings. This process of incrementally adding elements to the ratings proceeded until raters were applying to each practice essay all six elements on the Analytic scale. Approximately fifteen additional essays then were rated and discrepancies in ratings discussed. Rater comments highlighting differences between levels of quality within each element as well as differences between elements were summarized. (See Appendix G.)

A pilot test to determine the degree of rater agreement then was conducted. Each rater independently rated typed copies of Topic 1 and Topic 2 essays of fifteen high school students (total = 30 essays). Coefficients of reliability were calculated in the manner described under the section, "Impressionistic Rater Training." The rating procedure design for the Analytic scale training, as well as the final study Analytic ratings, is depicted in Figure 7.

The variance components from which the reliability coefficients were derived are presented in Appendix M . Table 3 presents the coefficients

for the total scale reliability and the six analytic subscales.

(Insert Table 3 about here)

As shown in Table 3, no reliability estimate was lower than .80. Consequently, the final study Analytic rating session was scheduled for the following day.

Analytic Scale Rating Session

Prior to rating the final study essays, one and one-half hours were devoted to reviewing the Analytic scale and applying it to approximately fifteen practice essays. Following the practice, each rater received a packet containing a complete set of the 224 typed essays, as described in the section, "Rating Procedures."

The Analytic scale ratings occurred over a two-day period. Each rater was encouraged to rate at her own pace and to take breaks in rating as needed. One rater completed all ratings within ten hours, including breaks, with approximately six hours on day one and four hours on day two. The second rater required approximately thirteen hours, including breaks, with six hours on the first day and seven hours on the second day.

A tally of the ratings was maintained by the researcher during the rating sessions. Most discrepancies of more than two points were discussed after the scores were recorded. A common set of papers periodically were read and rated. Characteristics of the common essays were discussed by the raters after the scores were recorded.

Using the method described in "Impressionistic Rater Scale Training," rater reliability of the 224 essays was calculated, Appendix N presents the estimates of variance components. Table 4 displays the rater reliability coefficients.

Table 3

RATER RELIABILITIES OF ANALYTIC SCALE PILOT TEST

Scale Element:	I Focus	II Devel-opment	III Organi-zation	IV Sup-port	V Para-graph-ing	VI Mech-anics	TOTAL SCALE
Reliability	.94	.92	.94	.83	.94	.80	.90

Table 4

RATER RELIABILITIES OF ANALYTIC SCALE FINAL STUDY

Scale Element	I Focus	II Devel-opment	III Organi-zation	IV Sup-port	V Para-graph-ing	VI Mech-anics	TOTAL SCALE
Reliability	.84	.80	.85	.85	.80	.95	.90

A comparison of the Analytic pilot test coefficients (Table 3) with final study coefficients revealed a decrease in reliability estimates with the exceptions of subscales II and VI, which increased in reliability estimates. The total scale reliability did not vary from that of the pilot test (.90).

Basic Measurement Properties of Analytic Scale

For each case, essay scores were calculated by summing the scores assigned by both raters to both essays for each subscale and across subscales. The range of possible points for each Analytic subscale thus was 4 through 16, with a total score (the sum of subscale scores) range of 24 through 96. Means and standard deviations for the study population and by class are displayed in Table 5.

(Insert Table 5 about here)

Table 5 indicates a high degree of consistency in mean scores across subscales for the total population (range = $\overline{X}$ = 9.9 - $\overline{X}$ = 10.7). A similar consistency across subscales is evident by class. Also across subscales, the means of Class V and Class II consistently ranked highest and the mean performances of Class I ranked lowest. A comparison of class mean performance on the Impressionistic ratings (Table 2) revealed the same pattern. The total score means of Class V ($\overline{X}$ = 78.8) and Class I ($\overline{X}$ = 67.2), rated by their respective teachers as above average in writing ability, were above the total population mean of 62.5. The greatest discrepancy between teacher estimates of performance and class mean scores was evidenced by Class I which was rated as "average" in writing ability yet consistently ranked lowest in performance.

Table 5

MEANS AND STANDARD DEVIATIONS ON ANALYTIC SCALES

ELEMENT	I Focus	II Devel-opment	III Organi-zation	IV Support	V Para-graph	VI Mech-anics	TOTAL SCALE
Range of Possible Scores	4-16	4-16	4-16	4-16	4-16	4-16	24-96
CLASS							
I. (n=10)	$\bar{X}$= 7.3 SD= .9	$\bar{X}$= 7.2 SD= 1.3	$\bar{X}$= 5.7 SD= 1.2	$\bar{X}$= 6.9 SD= 1.2	$\bar{X}$= 6.2 SD= 1.2	$\bar{X}$= 8.3 SD= 1.9	$\bar{X}$=41.6 SD= 4.5
II. (n=19)	$\bar{X}$=10.7 SD= 2.1	$\bar{X}$=11.2 SD= 2.2	$\bar{X}$=10.8 SD= 2.3	$\bar{X}$=11.2 SD= 2.4	$\bar{X}$=11.6 SD= 2.0	$\bar{X}$=11.7 SD= 2.3	$\bar{X}$=67.2 SD=11.4
III. (n=16)	$\bar{X}$= 9.3 SD= 3.3	$\bar{X}$= 9.3 SD= 3.1	$\bar{X}$= 8.8 SD= 2.9	$\bar{X}$= 8.8 SD= 1.7	$\bar{X}$= 9.6 SD= 2.6	$\bar{X}$= 9.0 SD= 2.3	$\bar{X}$=54.8 SD=14.1
IV. (n=19)	$\bar{X}$= 8.9 SD= 2.3	$\bar{X}$= 8.6 SD= 2.7	$\bar{X}$= 8.3 SD= 2.9	$\bar{X}$= 9.2 SD= 2.5	$\bar{X}$= 9.2 SD= 2.9	$\bar{X}$= 9.9 SD= 2.3	$\bar{X}$=54.1 SD=13.4
V. (n=32)	$\bar{X}$=13.3 SD= 2.5	$\bar{X}$=13.3 SD= 2.4	$\bar{X}$=12.8 SD= 2.4	$\bar{X}$=13.4 SD= 2.2	$\bar{X}$=13.5 SD= 2.1	$\bar{X}$=12.6 SD= 1.5	$\bar{X}$=78.8 SD=11.3
VI. (n=16)	$\bar{X}$=10.1 SD= 2.6	$\bar{X}$= 9.3 SD= 3.4	$\bar{X}$= 8.8 SD= 3.0	$\bar{X}$= 8.3 SD= 1.9	$\bar{X}$= 9.4 SD= 2.6	$\bar{X}$= 9.1 SD= 2.6	$\bar{X}$=54.9 SD=13.2
TOTAL (n=112)	$\bar{X}$=10.6 SD= 3.1	$\bar{X}$=10.7 SD= 3.4	$\bar{X}$= 9.9 SD= 3.4	$\bar{X}$=10.3 SD= 3.1	$\bar{X}$=10.6 SD= 3.2	$\bar{X}$=10.6 SD= 2.6	$\bar{X}$=62.5 SD=16.9

Objective Item Test Procedures

A thirty-seven item objective test consisting of six Subsections was administered to participating students. As described under the section, "The Objective Measure," six different forms of the test were assembled and randomly assigned to students within each class who, in turn, were randomly assigned to take the test on one of three measurement occasions. Refer to the section, "Study Procedures -- Overview," for a complete description of the administrative procedures. A total of 117 objective tests were collected and scored by assigning 1 to correct responses, 0 to incorrect responses. Items for which there was no response were not entered into the analyses. For purposes of analysis, items were consecutively numbered, 1 through 37. These numbers are indicated on the Appendix I copy of the measure.

Basic Measurement Properties of Objective Item Test

To obtain basic information on the measurement properties of the test, means, standard deviations, reliabilities and item inter-correlations were calculated. Means, standard deviations, and average percent correct on the total test and six Subsections for the total population are displayed in Table 6.

(Insert Tables 6 and 7 about here)

As indicated in Table 6, the average percent of items answered correctly on the total test was 76% ($\bar{X}$ = 28.39). The average percent correct for Subsection III, Supporting Detail (93%), and Subsection IV, Paragraphing (83%), were considerably above that of the total test, while the same statistic for Subsection IV, Paragraph Analysis (69%), was below the average.

Table 7 presents means and standard deviations by class. The pattern of class mean performance on the total test was similar to that of class mean performances on both Impressionistic and Analytic rating scores. Class

Table 6

MEANS AND STANDARD DEVIATIONS ON TOTAL OBJECTIVE TEST
AND SUBSECTIONS FOR THE TOTAL POPULATION

Subsection	Items	Means	SD	N
I (Main Idea)	1- 5	3.56 (71%)	1.13	113
II (Development)	6-10	3.54 (71%)	1.05	114
III (Supporting Detail)	11-15	4.63 (93%)	.88	114
IV (Paragraphing)	16-20	4.23 (85%)	1.09	115
V (Organization)	21-25	3.93 (79%)	1.38	116
VI (Paragraph Analysis)	26-37	8.21 (69%)	2.56	114
TOTAL TEST	1-37	28.39 (76%)	5.45	109

TABLE 7

MEANS AND STANDARD DEVIATIONS ON TOTAL OBJECTIVE TEST AND SUBSECTIONS BY CLASS

Sub-sections:	I (Main Idea)	II (Devel-opment)	III (Support-ing Detail)	IV (Para-graph-ing)	V (Organi-zation)	VI (Para-graph Analysis)	TOTAL TEST
Items:	1-5	6-10	11-15	16-20	21-25	26-37	1-37
CLASS							
I.	$\bar{X}$=2.83	3.0	4.25	3.08	3.17	6.18	22.10
(n=12)	SD .94	1.21	1.22	1.51	1.59	.76	5.56
II.	$\bar{X}$=3.43	4.0	4.86	4.71	4.10	8.70	29.70
(n=21)	SD= .98	.89	.36	.72	1.26	1.22	3.50
III.	$\bar{X}$=2.93	3.13	4.13	3.94	3.63	7.19	25.36
(n=15)	SD=1.44	.64	1.20	.68	1.41	2.43	4.09
IV.	$\bar{X}$=3.82	3.23	4.50	4.26	3.56	6.61	26.69
(n=17)	SD=1.19	1.39	.73	.90	1.73	2.64	6.16
V.	$\bar{X}$=3.97	4.0	5.0	4.77	4.74	10.58	33.07
(n=31)	SD= .91	.89	.0	.50	.58	1.18	1.86
VI.	$\bar{X}$=3.77	3.17	4.56	3.72	3.50	7.39	26.52
(n=17)	SD=1.03	.79	1.25	1.49	1.43	2.55	5.15

V ($\bar{X}$=33.07) and Class II ($\bar{X}$=29.70) were the two highest ranked, with Class I ($\bar{X}$=22.10) ranked lowest.

To estimate the degree of homogeneity of items within subsections, coefficient alpha was calculated based on mean scores for each Subsection with adjustments made for missing responses. Total test reliability was calculated at .83, while coefficients for subsections ranged from .71 (Subsection III, Supporting Detail) to .33 (Subsection II, Development), as shown in Table 8.

TABLE 8

OBJECTIVE MEASURE COEFFICIENTS OF RELIABILITY

Subsection:	I (Main Idea)	II (Development)	III (Supporting Detail)	IV (Paragraphing)	V (Organization)	VI (Paragraph Analysis)	TOTAL TEST
Reliability	.34	.33	.71	.57	.68	.69	.83
N	109	109	109	109	109	109	109

Separate analyses were performed on items within Subsection VI, which were intended to measure the following skill areas: main idea (items 26, 30, 34), development (items 27, 31, 35), organization (items 28, 32, 36), and supporting detail (items 29, 33, 37). Note that these skill areas are analogous to those covered in Subsection I (main idea), Subsection II (development), Subsection III (supporting detail), and Subsection V (organization) but differ in the stimulus characteristics of test item content. The results of these correlations are presented in Table 9.

(Insert Table 9 here)

TABLE 9

OBJECTIVE MEASURE SUBSECTION VI ITEM CORRELATIONS

I	II	III	IV
Subsection VI Items	Correlation with Analogous Sub-sections	Correlation with Similar Sub. VI Items	Correlation with Subsection VI Total
I (Main Idea) Items:	(with Sub. I, "Main Idea")	(with Items 26,30,34)	
26	.14	.67	.47
30	.10	.73	.42
34	.28	.66	.48
II (Development) Items:	(with Sub. II, "Development")	(with Items 27,31,35)	
27	.12	.68	.50
31	.20	.67	.45
35	.15	.56	.40
III (Organization)	(with Sub. V, "Organization")	(with Items 28,32,36)	
28	.10	.41	.34
32	.19	.78	.56
36	.41	.75	.59
IV (Supporting Detail)	(with Sub. III, "Supporting Detail")	(with Items 29,33,37)	
29	.20	.73	.53
33	.11	.71	.41
37	.12	.75	.59

Column I of Table 9 lists items within Subsection VI according to skill area. Column II presents correlations between the items and analogous Subsections. Column III presents correlations between each item and related items within Subsection VI. Correlations between each item and Subsection VI, when Subsection VI is treated as a twelve-item subtest, are displayed in Column IV.

A comparison of Columns II and III indicates that the relationship among items within Subsection VI is considerably stronger than the relationship of the same items to analogous Subsections. This discrepancy in the degree of association is understandable due to the differing requirements of items within Subsection VI and those of other Subsections. Items within Subsection VI required students to make discriminations among sentences within 75 to 100 word passages. In contrast, Subsections I through V involved discriminations at the level of the sentence.

An examination of Column IV reveals a relatively strong association between the items and Subsection VI. A comparison of Columns III and IV indicates that this relationship is less strong than that evidenced by similar items within Subsection VI. This difference is to be expected, as Subsection VI imposed four different response requirements (e.g., identifying the main idea, selecting statements needing additional support). When Subsection VI is classified into subtests comprised of similar item-types, a pattern of consistently strong item-subtest correlations emerges, as indicated in Column III.

The data in Table 9 suggest that items comprising Subsection VI can be treated as four separate subtests measuring skills analogous -- but not

necessarily identical -- to those covered in Subsections I, II, III and V. This preliminary analysis of items within the Objective measure resulted in the following classifications of items for further analyses.

Classification	Items
Main Idea	1- 5
Development	6-10
Supporting Detail	11-15
Paragraphing	16-20
Organization	21-25
Paragraph Analysis	26-37
Main Idea[1]	26,30,34
Development[1]	27,31,35
Organization[1]	28,32,36
Supporting Detail[1]	29,33,37

1=items within Subsection VI

FINDINGS

Findings--Overview

The study described in this report involved the development and application of three strategies to measure high school students' writing within the discourse domain of exposition. The study procedures were undertaken in order to obtain information on the relationships among scores derived from the three strategies. Specific research questions addressed by the study were as follows:

1. What is the relationship between Impressionistic and Analytic rating scale scores?
2. What is the relationship between Analytic scale scores and scale scores and those yielded by the Objective measure?
3. What is the relationship between Impressionistic scale scores and those yielded by the Objective measure?

To respond to these questions, correlational and regression analyses were performed on the data and are reported in this section. Relations between the two essay rating systems, Impressionistic and Analytic, will be described first. Associations between the two essay rating schemes and the Objective measure then will be reported, followed by a summary of major findings. Note that discussion of the findings, as well as their implications, are presented in the final section, Discussion.

Relationships Among Analytic and Impressionistic Scale Scores

Table 10 displays correlations among Impressionistic scale scores, the six Analytic subscales, and the Analytic scale total.

(Insert Table 10 about here)

TABLE 10

INTERCORRELATIONS AMONG IMPRESSIONISTIC AND ANALYTIC RATING SCALE SCORES

	Imp. Total	An. Total	I Focus	II Devel-opment	III Organ-ization	IV Sup-port	V Para-graph	VI Mech-anics
Impressionistic Total	1.00							
Analytic Total	.82	1.00						
Analytic Subscale:								
I. Focus	.65	.91	1.00					
II. Development	.69	.92	.90	1.00				
III. Organization	.76	.96	.89	.90	1.00			
IV. Support	.80	.86	.72	.71	.82	1.00		
V. Paragraphing	.74	.94	.80	.83	.91	.84	1.00	
VI. Mechanics	.79	.82	.69	.73	.72	.70	.70	1.00

A pattern of relationships among the six subscales comprising the Analytic scale is evident. The strength of the associations among the Analytic subscales is evidenced by the relatively high subscale-Analytic total correlations, which range from .96 (Organization) to .82 (Mechanics). As shown, the strongest set of relationships exists with Organization, which is highly related to Paragraphing (.91), Development (.90) and Focus (.89). Among the six subscales, Mechanics yielded the lowest range of intercorrelations (.69 to .73). These data suggest a trend of consistency in the ratings of writing performance across the categories measured by the Analytic scale.

Table 10 also provides correlations between Impressionistic scores and those yielded by the Analytic rating scale. The correlation between the Impressionistic scores and Analytic total scores (.81) indicates a strong relationship between the two rating strategies. This association extends to relationships between the six Analytic subscales and Impressionistic ratings. These correlations range from .65 (Focus) to .80 (Support).

To examine further the predictive relationship between Impressionistic and Analytic scores by statistically controlling the influence of the subscales on each other, Impressionistic scores were regressed on the six Analytic subscales. The six Analytic subscales accounted for approximately 75% of the variation ($F = 53.736$, $df = 6,105$, $p < .01$) in Impressionistic scores. Both unstandardized and standardized regression coefficients for the Analytic subscales are presented in Table 11.

(Insert Table 11 here)

Two subscales, Mechanics ($F=30.789$) and Support ($F=18.36$), proved to be statistically significant predictors of Impressionistic scale scores in the

model with all six subscales.

The relatively high degree of multicolinearity of the Analytic subscales (See Table 10) suggested that the relative importance of the subscales as predictors may be masked in the regression analyses. To examine this, a new variable was entered into the equation, Structure, which was comprised of the sum of scores on four subscales, Organization, Focus, Paragraphing, and Development. These subscales correlated strongly with the Analytic total and to each other, suggesting that it may be appropriate to treat them as a unitary variable (See Table 10). The results of this regression, which were significant ($F = 105.159$, $df = 3,108$, $p < .01$), are presented in Table 12.

(Insert Table 12 here)

The combination of the three subscales, Mechanics, Support, and Structure, account for approximately 74% of the variation in Impressionistic scale scores. Again, Support and Mechanics emerged as the strongest predictors to Impressionistic scores. These data suggest that ratings on elements comprising the Structure variable are confounded with ratings on Mechanics and Support. When the influence of Mechanics and Support are statistically controlled for, the unique contribution of Structure as a rating category, is diminished.

Relationships Among Rating Scale Scores and the Objective Measure

Intercorrelations among the Impressionistic rating scores, Analytic total scores, and scores on the Objective measure are presented in Table 13. These data reveal similar patterns of correlation between Objective

TABLE 11

REGRESSION OF IMPRESSIONISTIC SCALE SCORES ON SIX ANALYTIC SUBSCALES

Analytic Subscales	Unstandardized Coefficient	Standard Error of β	Standardized Coefficient	F
Mechanics	.585	.106	.424	30.789*
Focus	.233	.147	.196	2.537
Development	.046	.160	.040	.082
Organization	.293	.182	.265	2.492
Support	.514	.120	.424	18.365*
Paragraphing	.033	.148	.028	

df = 1, 105

* = $p < .01$

$r^2 = .75$

R = .87

TABLE 12

REGRESSION OF IMPRESSIONISTIC SCALE SCORES ON THREE ANALYTIC SUBSCALES

Analytic Subscales	Unstandardized Coefficient	Standard Error of β	Standardized Coefficient	F
Mechanics	.582	.104	.422	31.468*
Support	.560	.105	.462	28.502*
Structure[1]	.017	.028	.056	.370

df = 1, 108

* = $p < .01$

Structure[1] = Composite score of Organization, Focus, Development, Paragraphing

$r^2 = .74$

R = .86

measure scores and both Analytic and Impressionistic rating scale scores. This similarity is to be expected given the relatively high degree of association (.81) between the two rating systems. The correlation between the Objective measure total and Analytic scale total was .61, while the correlation with Impressionistic scores was .65. As shown, Paragraphing (Subtest IV) and Paragraph Analysis (Subtest VI) emerged as the Objective subtests most highly related to the two rating scale strategies. Note that Paragraph Analysis is comprised of the twelve items constituting subtests A through D.

The Impressionistic and Analytic total scales were independently regressed on the five major subtests listed in Table 13. Regression coefficients yielded by these calculations are reported in Tables 14 and Table 15.

(Insert Tables 13, 14, and 15 here)

Results of the regression analyses for the Impressionistic scale ($F = 12.853$, $df = 6, 92$, $p < .01$) and for the Analytic scale ($F = 10.338$, $df = 6, 92$, $p < .01$) were statistically significant. Tables 14 and 15 indicate that the five Objective measure subtests jointly account for approximately 46% of variation in Impressionistic scale scores and for approximately 40% of the variation in Analytic scale scores. Two Objective subtests, Paragraphing and Paragraph Analysis, proved to be the strongest predictors to both Impressionistic and Analytic scale total scores.

Recall that an area of inquiry for this study concerned the relationships between Objective measure subtests and analogous subscales of the Analytic rating scale. To examine these relationships, a series of regression analyses were performed. In each analysis, an Analytic subscale was regressed on analogous Objective subtests. Results of these analyses are displayed in

TABLE 13'

CORRELATIONS AMONG ANALYTIC AND IMPRESSIONISTIC TOTAL SCORES AND OBJECTIVE MEASURE SCORES

	Imp. Total	An. Total	I Main Idea	II Devel-opment	III Supt. Detail	IV Para-graph	V Organ-ization	VI Para. Anal-ysis	A Main Idea[1]	B Devel-opment[1]	C Organ-ization[1]	D Supt. Detail[1]	Total Test
Impressionistic Total	1.00												
Analytic Total	.81	1.00											
Objective Subtests													
I. Main Idea	.24	.23	1.00										
II. Development	.46	.37	.17	1.00									
III. Supt. Detail	.35	.35	.11	.35	1.00								
IV. Paragraph	.55	.49	.29	.43	.37	1.00							
V. Organization	.45	.42	.26	.38	.32	.39	1.00						
VI. Para. Analysis	.53	.53	.36	.47	.35	.43	.43	1.00					
A. Main Idea[1]	.30	.35	.24	.36	.12	.23	.12	.72	1.00				
B. Develop-ment[1]	.40	.40	.26	.33	.30	.35	.37	.74	.40	1.00			
C. Organization[1]	.39	.33	.32	.35	.35	.32	.33	.75	.44	.34	1.00		
D. Supt. Detail[1]	.47	.51	.21	.33	.24	.33	.42	.71	.30	.38	.44	1.00	
Objective Total Test	.65	.61	.53	.66	.55	.68	.68	.85	.54	.66	.67	.63	1.00

1 = Items within Paragraph Analysis

TABLE 14

REGRESSION OF ANALYTIC SCALE TOTAL ON SIX OBJECTIVE MEASURE SUBTESTS

Objective Measure Subtests	Unstandardized Coefficient	Standard Error of β	Standardized Coefficient	F
Main Idea	.141	1.300	.009	.012
Development	.428	1.595	.026	.072
Supporting Detail	1.502	1.669	.082	.811
Paragraphing	4.037	1.514	.260	7.105*
Organization	1.771	1.173	.142	2.277
Paragraph Analysis	2.184	.674	.330	10.488*

df = 1, 92

* = $p < .01$

R = .66

r^2 = .44

TABLE 15

REGRESSION OF IMPRESSIONISTIC SCALE TOTAL ON SIX OBJECTIVE MEASURE SUBTESTS

Objective Measure Subtests	Unstandardized Coefficient	Standard Error of β	Standardized Coefficient	F
Main Idea	.025	.274	.007	.008
Development	.497	.336	.137	2.182
Supporting Detail	.209	.352	.051	.354
Paragraphing	1.06	.319	.310	11.088*
Organization	.421	.248	.153	2.894
Paragraph Analysis	.371	.142	.254	6.08*

df = 1, 92

* = $p < .01$

R = .68

r^2 = .46

Table 16.

(Insert Table 16 here)

The analyses reported in Table 16 resulted in significant F ratios for each Objective subtest, with the exception of Organization[1]. The amount of variation in Analytic subscale scores accounted for by analogous Objective subtests ranged from 14% to 22%. The correlations are only slightly less than the correlation between the five subtests operating jointly and the Analytic scale total (44%).

Summary of Findings

Correlational and regression analyses performed on the data of the study resulted in the following major findings:

1. Scores yielded by the Impressionistic and Analytic rating strategies were highly related.
    a. The six subscales comprising the Analytic rating scale were highly intercorrelated.
    b. Of the six Analytic subscales, Mechanics and Support proved to be most strongly associated with Impressionistic rating scores.
2. The Objective measure total score proved to be moderately related to both the Analytic and Impressionistic total scale scores.
    a. Of the six Objective measure subtests, Paragraphing and Paragraph Analysis proved to be most strongly related to both Impressionistic and Analytic total scale scores.
    b. A moderately strong relationship existed between Analytic subscales and analogous Objective measure subtests.

TABLE 16

REGRESSION OF ANALYTIC SUBSCALES ON ANALOGOUS OBJECTIVE MEASURE SUBTESTS

Analytic Subscale	Objective Subtest	Unstandardized Coefficients	Standard Error of β	Standardized Coefficients	F	df	R	r^2
I. Focus	Main Idea	.518	.269	.188	3.709*	1,96	.37	.14
	Main Idea[1]	1.051	.375	.273	7.826*	1,96		
II. Development	Development	.660	.307	.213	4.619*	1,96	.40	.17
	Development[1]	.953	.333	.283	8.171*	1,96		
III. Support	Supporting Detail	.545	.313	.161	3.026*	1,96	.47	.22
	Supporting Detail[1]	1.548	.347	.412	19.896*			
IV. Paragraphing	Paragraphing	1.303	.269	.442	23.511*	1,97	.44	.20
V. Organization	Organization	.905	.360	.360	13.516*	1,96	.42	.18
	Organization[1]	.514	.131	.131	1.797	1,96		

1 = items within Subtest VI, Paragraph Analysis

* = $p < .01$

c. Item content (e.g., stimulus and response attributes) appeared to exert an influence on the predictive relationship between the Objective measure and the two rating strategies.

DISCUSSION

Overview

It is generally acknowledged that the measurement of writing and writing-related competencies presents unique technical and validation problems. From pioneer assessment efforts (Starch and Elliot, 1912) to current composition tests (Haugh and Brown, 1969), measurement experts have contended with such recurring technical problems as fluctuating test reliability, implementation of efficient scoring procedures, and examining the validity of indirect, i.e., objective, measures of writing (Godshalk, Swineford, and Coffman 1966).

In recent years, additional requirements for measures of writing have arisen along with the demand for competency-based tests of basic skills, including writing. These requirements call for measures that not only are technically sound, but which also serve as meaningful, efficient indicators of clearly defined writing competencies. Moreover, teachers, educational administrators, legislators and parents are seeking measures that carry readily interpretable implications for instructional diagnosis and remediation.

The present study thus was undertaken as an effort to address some of the measurement concerns brought to light by the competency-based assessment movement. The primary aim of the study was to examine relationships among three writing assessment strategies which are potentially useful within the context of competency-based testing. Two of the strategies manipulated in the study can be classified as direct measures of performance, that is, they involved collecting and rating actual samples of student writing. The third strategy, an indirect measure of performance, involved administration

of an objective test of writing skills presumably linked to the two direct measure strategies.

The first section of this discussion will address the findings pertaining to the relationships among the three writing assessment strategies. Implications of the findings for both instructional and measurement practice then will be presented.

Discussion of Findings

Relationships Among the Direct Measures of Writing

The chief distinction between the two direct measures of writing employed in the present study lay in the form, and underlying assumptions, of the criteria by which the students' writing samples were judged. One of the direct measure strategies involved application of an Analytic rating scale in which separate scores were assigned to each of six different characteristics of the writing products. Analytic scoring procedures, such as the one employed, presume that a piece of writing can be viewed as consisting of component, but not necessarily independent, parts which are worthy of individual scrutiny. Analytic rating procedures aptly meet the requirements of competency-based assessment programs which call for unambiguous information on specific skill deficits and strengths.

In contrast, impressionistic rating procedures assume that "...each factor that makes up writing skill is related to all other factors and that one factor cannot be easily separated from others" (Office of the Los Angeles County Superintendent of Schools, 1977). In keeping with this assumption, the Impressionistic rating procedure applied in the study required raters to make a judgment regarding the overall quality of each writing sample as an example

of effective exposition. When large aggregates of students are to be tested, as in the case of many competency-based assessment programs, Impressionistic rating procedures can prove to be an efficient direct measure of writing performance.

As reported, findings of the study revealed a strong degree of association among the scores yielded by the Analytic subscales, Analytic total score and Impressionistic scale scores. On the basis of these data alone, it is tempting to infer that the Analytic subscales actually tapped a single unitary dimension of writing comparable to that measured by the Impressionistic rating scale. Such an interpretation, however, overlooks an important facet of the entire measurement process -- the writing stimulus, e.g., the writing topics and accompanying directions.

Recall that the writing topics and directions employed in the study were specifically designed to elicit writing developed through a logically arranged structure of generalizations supported by specifics. Directions accompanying the topic prompted students in the following way:

> Remember, the purpose of your essay is to give an informative explanation... Back up your ideas with specific support, such as examples, facts, and other details. Make sure your essay is well-organized.

As anticipated, the writing stimuli promoted homogeneity in the rhetorical structure of the majority of the students' writing samples, of course, with varying degrees of quality in execution. It also is important to remember that the Analytic scale purposefully excluded such elusive qualities as "fluency of language" or "maturity of thought" which are likely to influence reader response but which are tenuously amenable to instructional manipulation. Consequently, Analytic raters were expected to attend exclusively to structural and mechanical characteristics of the writing samples. It is thus

possible that the relative conformity of rhetorical structures exhibited by the student essays accounted for the correlations among scores across the subscales. Whether differential subscale scores will emerge when writing samples display more varied structural patterns is an area worthy of additional inquiry.

It is also possible that the writing stimuli bore an indirect influence on the scores assigned by Impressionistic scale raters. Once Impressionistic raters became habituated to the structural patterns of the majority of writing samples (through practice and training), they may have attended to a few prominent features which most noticeably discriminated among the essays. Impressionistic ratings on the overall quality of the writing samples may, in fact, reflect rater judgments on a select number of highly discriminating features.

On the basis of regression analyses in which Impressionistic scale scores were regressed on the six Analytic subscales, one could infer that the most highly discriminating features of the essays were mechanical quality (e.g., Analytic subscale "Mechanics") and the inclusion of statements to support generalizations (e.g., Analytic subscale "Support"). These regression analyses were undertaken to examine the unique contribution of each Analytic subscale as a predictor to Impressionistic scores. When a new variable, Structure, which was a composite score of subscales Focus, Organization, Paragraphing and Development, was entered into the equation, Mechanics and Support again proved to be most highly related to Impressionistic scale scores.

Conceptually, the emergence of the two subscales Mechanics and Support as predictors to Impressionistic ratings is understandable. Judgments re-

garding a writer's command over mechanical aspects of writing can be made independently of the overall structure of a written product. Similarly, judgments regarding the adequacy of support for generalizations embedded within an essay can be made without reference to the way in which the dominant main idea, if present, is developed, organized and arranged into paragraphs. Analytic and Impressionistic raters may have attended to two readily discernible, and thus discriminating, features of the writing samples: Mechanics and Support.

Findings on the relationships among the Impressionistic and Analytic rating procedures employed in the study open some provocative lines of inquiry concerning direct measures of writing. For example, will writing stimuli which are less strongly prompted elicit a range of rhetorical structures in students' expository writing? If so, will the Analytic subscales developed for the study yield differential scores according to the characteristics and quality of varying structural patterns? An analog to such inquiries would involve application of Impressionistic rating procedures to Analytically scored essays. A departure from the present methodology, and one similar to that employed by Diederich (1961), would require Impressionistic raters to record the features that most strongly influenced their judgments regarding overall quality of the samples. The extent to which Analytic subscales predict to Impressionistic features may corroborate, or extend, the findings of the present study.

Relationships Between the Indirect and Direct Measures of Writing

Contrasting with both the Analytic and Impressionistic rating procedures was the indirect measure of writing employed in the study, a thirty-seven item Objective test. That objective tests of writing are at best indirect

predictors of writing competence is clearly evident. Nonetheless, given the efficiency with which objective measures can be adminstered and scored, they remain a potentially useful measurement strategy, especially for large scale competency-based assessment programs. Several correlational studies (e.g., Godshalk, 1966) have demonstrated relatively strong relationships between performance on objective tests of writing and actual production of writing samples. These correlational studies indicate that various constellations of skills measured by objective items are related in some unspecified manner to actual writing performance. If objective strategies are to be meaningful indicators of specific writing competencies, though, additional information regarding relationships between well-defined classes of objective tasks and equally well-defined writing production tasks is needed. In an effort to address this area of inquiry, objective items employed in the present study were derived from a set of explicit item generation guidelines. The development strategy permitted examination of relationships among classes of objective items with reference to their predictive relationship to clearly defined writing tasks.

The findings from the study revealed a moderately positive relationship between scores derived from the Objective strategy and scores yielded by the two essay-rating strategies. The positive relationship was expected, as the objective items were designed to assess, at the levels of recognition and discrimination, those categories of skill measure by the rating strategies at the level of production.

An examination of the properties of the Objective test items provided an interesting insight on the relationship between Objective measure scores and essay-rating scores. The results from item intercorrelations pointed

to the possible influence of item content (e.g., stimulus and response attributes) on test performance which, in turn, was likely to have affected the predictive value of the items. For example, the possible influence of item content was evidenced by the set of items comprising the subtest Paragraph Analysis. The Paragraph Analysis subtest contained 75 to 100 word passages, drawn from actual samples of student writing, which were followed by a series of four different questions. The skill areas tapped by the Paragraph Analysis questions were analogous to those covered in other subtests of the Objective measure, but differed in the stimulus characteristics of test item content. Interestingly, the relationships among items within the Paragraph Analysis subtest were considerably stronger than the relationship of the same items to analogous subtests.

The trend of findings derived from results of item correlations underscores the importance of method -- in this case, item content -- in the estimation of skill acquisition. Especially worthy of further examination are properties of items within the subtests Paragraphing and Paragraph Analysis, which proved to be the strongest predictors to essay-rating total scores.

Both Paragraphing and Paragraph Analysis subtests required students to make discriminations among statements within prose passages. Given that these measurement tasks required a certain level of reading comprehension, it is likely that reading ability intervened in the relationship between performance on the subtests and essay-rating scores. Interestingly, other studies (e.g., Godshalk, 1969) reporting strong relationships between indirect and direct writing measures included objective items containing complex reading passages.

Hypotheses concerning the influence of reading ability on test item prop-

erties could, of course, be tested by statistically controlling for students' level of reading. A more compelling line of investigation, however, would involve a priori sampling from populations distinctly different in reading competence, e.g., both high and low reading levels. A counterpart to the investigation would involve examination of relationships between classes of objective items and judgments of the groups' writing competence. An investigation of this nature would augment the current data base on the influence of item content and, concomitantly, reading ability, in objective measures of writing-related competencies.

Implications for Educational Practice

Procedures of the study, and resultant findings, yield interesting -- and direct -- implications for both the measurement and teaching of writing competencies. Perhaps one of the most striking findings to emerge concerns the strong relationship between the two direct measures of writing, the Analytic and Impressionistic rating scales. Given a writing topic and directions comparable to those employed in the study, the use of Impressionistic rating procedures would appear to be the most efficient means for assigning scores to collected samples of students' writing. A benefit of Impressionistic rating procedures is that they permit raters to respond intuitively to the overall quality of a piece of writing, free from any arbitrary bias that might be introduced by analytic rating strategies. As a corollary, though, Impressionistic rating strategies also preclude public disclosure of the criteria of judgment. This latter characteristic violates a fundamental tenet of many competency-based assessment programs which call for readily interpretable information on specific skill strengths and deficits. A potential

resolution to this dilemma would involve application of the Analytic scale as a rubric for training raters. Such a procedure would have the advantage of alerting raters to prominent characteristics of writing within the discourse domain of exposition, thereby efficiently facilitating reliability of ratings. Application of the Analytic scale as a training rubric would have the additional advantage of disclosing the criteria of rater judgment.

While Impressionistic rating scales may prove to be efficient instruments, especially for large scale assessments, it is likely that the Analytic scale could serve as a useful evaluative and instructional tool at the classroom level. For example, teachers could employ the Analytic subscales for planning instruction in expository discourse and for formative evaluation of students' writing. The Analytic scale could thus be used to satisfy the requirements for instructional remediation demanded by such legislation as the California Hart Bill. Naturally, the diagnostic value of the Analytic scale should be further examined by applying it to writing samples of heterogeneous groups of students and by varying the writing stimuli.

In contrast to the apparent utility of direct measures of writing, the contribution of objective items, such as those employed in the present study, requires further examination. In this study, performance on the Objective measure was relatively consistent and high across the subtests. Yet performance on the classes of items, with the exception of two subtests, proved to be only moderately related to scores yielded by the Analytic and Impressionistic rating procedures. Clearly, if indirect measures are to be used as a means for measuring specific writing-related competencies, additional information regarding the properties of the items -- as well as their relation to direct measures of writing -- is needed.

Findings concerning the Objective measure do carry some implications for instruction in composition. As previously described, items included in the Objective measure were largely derived from a review of selected-response exercises contained in secondary composition textbooks and workbooks. The moderate correlation between the objective items employed in the study and the essay-rating scores sheds some doubt on the efficacy of employing selected-response practice as the primary mechanism for promoting attainment of skills in expository writing.

To conclude, the findings of the present study suggest directions for research in the relatively uncharted area of competency-based assessment of writing. Additionally, study procedures, and the assumptions underlying them, carry implications for current measurement and instructional practice in composition.

References

Braddock, R., Lloyd-Jones, R., & Schoer, L. Research in written composition. Champaign, Illinois: National Council of Teachers of English, 1963.

Brelend, H. M. A study of college English placement and the test of standard written English. Princeton, New Jersey: Educational Testing Service, January, 1977.

Brooks, C., Warren, R. Modern rhetoric, shorter edition. New York: Harcourt, Brace and World, Inc., 1961.

Coffman, W. E. Essay exams. In R. L. Thorndike (Ed.), Educational Measurement, Second Edition. Washington, D. C.: American Council on Education, 1971.

Cooper, C. Holistic evaluation of writing. In C. Cooper, & L. Odell (Eds.), Evaluating writing: Describing, measuring, judging. Buffalo: State University of New York at Buffalo, 1977.

Cronbach, L., Gleser, G., Nanda, H., & Rajartnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. Canada: John Wiley and Sons, 1972.

Diederich, P., French, J. W., & Carlton, S. Factors in the judgments of writing ability. Princeton, New Jersey: Educational Testing Service Research Bulletin, 1961.

Diederich, P. B. Measuring growth in English. National Council of Teachers of English, 1974.

Edgeworth, F. V. The element of chance in competitive examination. Journal of the Royal Statistical Society, 1890, LIII.

Eeley, E. G. The test satisfies an educational need. College Board Review, 1955, 25.

Eels, W. C. Reliability of repeated grading of essay-type examinations. Journal of Educational Psychology, 1930, 21.

Follman, J. C., & Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. Research in the teaching of English, 1967, 1.

French, J. W. Schools of thought in judging excellence of english themes. 1961 Proceedings of Invitational Conference on Testing Problems, Educational Testing Service, Princeton, New Jersey, 1962.

Gallo, J., & Rink, H. Shaping college writing, paragraph and essay. New York: Harcourt, Brace and World, 1968.

Glaser, R., & Resnick, L. Instructional Psychology. Annual Review of Psychology, 1972, 23.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Gorrell, R. M. Freshman Composition. In J. C. Gerber, et al., (Eds.), The College Teaching of English. New York: Appleton, Century, Crofts, 1965.

Grommon, A. H. Reviews of selected published tests in English. Urbana: Illinois: National Council of Teachers of English, 1976. ED 120 732.

Haugh, O. M., & Brown, J. K. College English Placement test. Boston: Houghton Mifflin, 1969.

Kincaid, G. L. Some factors affecting quality of students' writing. Unpublished ED. D. dissertation. (Michigan State University) 1953. Cited in R. Braddock. Research in written composition, Ill. National Council of Teacher's of English, 1963.

Klein, S. P., & Hart, F. Change and systematic factors affecting essay grades. Journal of Educational Measurement, 5, 1968.

Lloyd-Jones, R. Primary trait scoring. In C. Cooper, L. Odell (Eds.) Evaluating writing: describing, measuring, judging. Buffalo: State University of New York at Buffalo, 1977.

McColly, W. What does educational research say about the judging of writing ability? The Journal of Educational Research, 64, No. 4, December 1970.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) Evaluation in education: current applications. Berkeley, California: McCurchan 1974.

Office of the Los Angeles County Superintendent of Schools. A common ground for assessing competencies in written expression, review copy. Los Angeles: Division of Curriculum and Instructional Services, 1977.

Page, E. B. The analysis of essays by computer. Final Report, U. S. Office of Education Project 6-1318, The University of Connecticut, Storrs, 1969.

Remondino, C. A. Factorial analysis of the evaluation of scholastic compositions in the mother tongue. British Journal of Educational Psychology, 1959, 30.

Snidman, N. S., & Quellmalz, E. S. Issues in criterion-referenced test construction. Paper presented at the annual meeting of the California Educational Research Association, San Diego, California, November, 1976.

Stalnaker, J. The construction and results of a twelve-hour test in English composition. School and Society, 39, 1934.

Stanley, J. C. Analysis of variance principles applied to the grading essay tests. Journal of Experimental Education. 1962, 30 (3).

Starch, D., & Elliott, E. C. Reliability of the grading of high school work in English. School Review, 1912, 20.

Veal, L. R. & Tillman, M. Mode of discourse variation in the evaluation of children's writing. Research in the Teaching of English, 1971, 5: 37-45.

Warriner, J. & Griffith, F. English grammar and composition, complete course. New York: Harcourt, Brace & Jovanovich, 1973.

West, W. Developing writing skills, second edition. New Jersey: Prentice-Hall, 1973.

Yeasmen, N., & Barker, D. A. Half a century of research on essay testing. Improving College and University Teaching, 1973, XXI(1).